BAYESIAN MODELS

FOR DNA MICROARRAY DATA ANALYSIS

A Dissertation

by

KYEONG EUN LEE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2004

Major Subject: Statistics

BAYESIAN MODELS

FOR DNA MICROARRAY DATA ANALYSIS

A Dissertation

by

KYEONG EUN LEE

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Bani K. Mallick (Co-Chair of Committee)	James A. Calvin (Co-Chair of Committee)
Naisyin Wang (Member)	Edward R. Dougherty (Member)

James A. Calvin
(Head of Department)

May 2004

Major Subject: Statistics

ABSTRACT

Bayesian Models

for DNA Microarray Data Analysis. (May 2004)

Kyeong Eun Lee, B.A., Kyungpook National University, Korea;

M.A., Seoul National University, Korea

Co–Chairs of Advisory Committee: Dr. Bani K. Mallick
Dr. James A. Calvin

Selection of significant genes via expression patterns is important in a microarray problem. Owing to small sample size and large number of variables (genes), the selection process can be unstable. This research proposes a hierarchical Bayesian model for gene (variable) selection. We employ latent variables in a regression setting and use a Bayesian mixture prior to perform the variable selection. Due to the binary nature of the data, the posterior distributions of the parameters are not in explicit form, and we need to use a combination of truncated sampling and Markov Chain Monte Carlo (MCMC) based computation techniques to simulate the posterior distributions. The Bayesian model is flexible enough to identify the significant genes as well as to perform future predictions. The method is applied to cancer classification via cDNA microarrays. In particular, the genes BRCA1 and BRCA2 are associated with a hereditary disposition to breast cancer, and the method is used to identify the set of significant genes to classify BRCA1 and others.

Microarray data can also be applied to survival models. We address the issue of how to reduce the dimension in building model by selecting significant genes as well as assessing the estimated survival curves. Additionally, we consider the well-

known Weibull regression and semiparametric proportional hazards (PH) models for survival analysis. With microarray data, we need to consider the case where the number of covariates p exceeds the number of samples n. Specifically, for a given vector of response values, which are times to event (death or censored times) and p gene expressions (covariates), we address the issue of how to reduce the dimension by selecting the responsible genes, which are controlling the survival time. This approach enables us to estimate the survival curve when $n << p$. In our approach, rather than fixing the number of selected genes, we will assign a prior distribution to this number. The approach creates additional flexibility by allowing the imposition of constraints, such as bounding the dimension via a prior, which in effect works as a penalty. To implement our methodology, we use a Markov Chain Monte Carlo (MCMC) method. We demonstrate the use of the methodology with (a) diffuse large B–cell lymphoma (DLBCL) complementary DNA (cDNA) data and (b) Breast Carcinoma data.

Lastly, we propose a mixture of Dirichlet process models using discrete wavelet transform for a curve clustering. In order to characterize these time–course gene expresssions, we consider them as trajectory functions of time and gene–specific parameters and obtain their wavelet coefficients by a discrete wavelet transform. We then build cluster curves using a mixture of Dirichlet process priors.

To my Lord Jesus, my parents, and my husband Ki Tak

ACKNOWLEDGMENTS

I would like to give my utmost praise to God and the Lord Jesus Christ who guided me to meet my adviser Dr. Bani K. Mallick and complete my degree. I would like to express my sincere gratitude and respect to him, for his intellectual advice and steadfast encouragement. I also want to give thanks to co-chair, Dr. James A. Calvin, and my dissertation committee members, Dr. Naisyin Wang and Dr. Edward R. Dougherty, for their support and understanding. I would like to express my gratitude to Dr. Longnecker and Dr. P. Fred Dahm. I have benefited from working with my close colleague, Shubhanka Ray, and I would like to thank him for his programming skills in curve clustering and our many discussions. I appreciate many colleagues in my department, especially Jeesun Jung and Sujung Choi, for their steadfast encouragement and friendship. I wish to express my gratitude to my parents and my parents-in-law for their encouragement. Finally, I would like to thank my husband Tak for his love and support.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

1.1. Background: cDNA Microarray

After the invention of the Southern blot, a method for searching for a specific DNA molecule which *introduced a one to one correspondence between clones and hybridization signals*, the use of non-porous solid supports and development of methods for high-density spatial synthesis of oligonucleotides opened up the world of DNA microarray technologies (Lander, 1999) which can provide expression measurements for thousands of genes at once (Duggan, 1999; Schena *et al.*, 1995). Two main types of microarray technologies are cDNA microarray and high density oligonucleotide array.

cDNA micoarray experimental procedures can be summarized as follows (Freind and Stoughton, 2002) :

1. Construct a microarray, containing a single-stranded DNA representing thousands of different genes. Each gene is assigned to a specific spot on a microarray which has thousands to millions of copies of a DNA strand.

2. Obtain mRNAs from two samples, one reference cell and one test cell.

3. Reverse transcribe unstable mRNA molecules into stable cDNAs and label with fluorescent dyes, red to cDNAs from the test samples and green to cDNAs from the reference samples.

4. Apply the labeled cDNA mixture solution to the microarray where it undergoes competitive binding.

The format and style of this dissertation follows that of *Biometrics*.

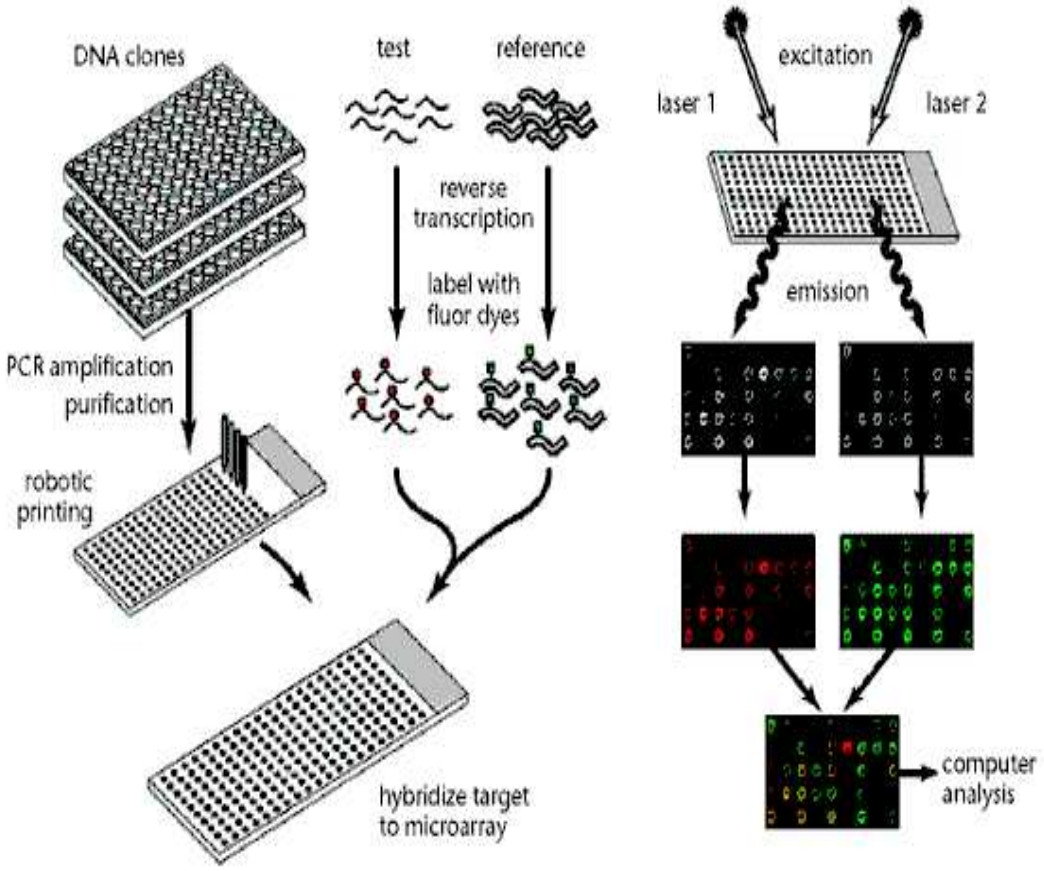

Fig. 1. cDNA Microarray Schema (Duggan *et al.*, 1999)

5. Using a scanner and a computer, obtain the ratio of the intensities of the red to green signals at each spot though image analysis.

More details on cDNA microarray are in Nguyen *et al.* (2002). Figure 1 provides a schematic overview of a cDNA microarray experiment.

1.2. Outline of Dissertation

Microarray problems can be classified as unsupervised, when only the expression data are available, and supervised, when a response measurement is available for each sample. In unsupervised problems the goal is mainly to identify distinct sets of genes with similar expressions, suggesting that they may be biologically related. Supervised and unsupervised problems also focus on finding sets of genes that, for example, relate to different kind of diseases, so that future tissue samples can be correctly classified. Traditional statistical methods for clustering and classification have been extensively applied to microarray data, see Eisen *et al.* (1998), Alizadeh *et al.* (2000) for clustering and Golub *et al.* (1999) and Hedenkalk *et al.* (2001) for classification. In the supervised case, a Bayesian approach to dimension reduction with a probit model has been applied by West *et al.* (2000) where rather than selecting actual genes, the singular-value decomposition is applied to the design matrix to reduce the dimension of the problem. In Chapter II, we mainly want to identify (select) important genes which are significantly more influential than the others for the classification process.

Often, the number of selected genes could be as large as 500 genes or even 2000 genes (Khan *et al.*, 1998; Alon *et al.*, 1999). Even in the studies which obtained relatively smaller numbers of genes, the numbers also to 50 to 100 (Golub *et al.*, 1999; Khan *et al.*, 2001). The main problem is that there is a very large set of genes and typically a small number of microarrays (sample points). So, selecting a large number genes is usually not advisable due to this small sample size as it can create an unreliable selection process. Dudoit *et al.* (2000) proposed a method for the identification of singly differentially expressed genes, considering a univariate testing problem for each gene and then corrected for multiple testing using adjusted p-values.

Tusher *et al.* (2001) created a Significance Analysis of Microarray (SAM) techinique, that assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance. Hastie *et al.* (2000) suggested gene shaving, a new class of clustering methods which tries to identify subsets of genes with coherent expression patterns and large variation across conditions. Kim *et al.* (2001) suggested a feature selection technique, by fixing the dimension of the selected genes and using a criterion based method to get the best subset of that dimension.

We suggest a model based approach to this variable selection problem. Rather than fixing the dimension (the number of selected genes of the problem), we assign a prior distribution over it. This creates additional flexibility as well as the ability to impose a constraint by limiting the dimension. So, the prior works as a penalty to create this constraint. We use a Markov Chain Monte Carlo (MCMC) (Gilks *et al.*, 1996) based stochastic search algorithm to identify the important genes. Though the model space is very large, as with p genes (say p=3000) we have 2^p models, and exhaustive computation over this model space is not possible. MCMC based stochastic search algorithms, which are less greedy and more efficient than most of the other alternatives, are successfully implemented to identify significant genes. We consider the model for binary events only, but the extension to multi-category data is straightforward.

In Chapter III, we consider a situation when survival times of (for example) cancer patients are of interest. In this setting, it is of interest to identify significant genes that might control the survival time of the patients. We also want to estimate the patient survival probabilities controlling for other covariates such as levels of clinical risk. For example, through gene expression profiling, Alizadel et al. (2000)

identified two distinct molecular subtypes of diffuse large B-cell lymphoma (DLBCL). Estimates of patient survival probabilities for the two groups were then compared using a Kaplan-Meier survival curve.

We suggest a gene selection technique using a Bayesian model based variable selection approach. Typical Bayesian variable selection methods are based on the assumptions of Gaussian distributions for the likelihood and use of conjugate mixture priors to obtain marginal distributions (George and McCulloch, 1993). We extend our models to the survival data context where the responses are time to event. We address the issue of how to select the significant genes as well as assess the survival curves using the Weibull regression or proportional hazards model (PH) where the sample size is much smaller than the number of variables (genes).

Current techniques used for variable selection in survival models include asymptotic procedures based on score tests and other approximate chi-square procedures. These asymptotic methods may not be valid here because n is much smaller than p. The literature on Bayesian variable selection for survival models is still sparse. Raftery, Madigan and Volinsky (1995) had used the approximate BIC criterion for variable selection. Ibrahim and Chen (2000) advocate a predictive approach which is efficient with small number of predictors (p is small). As our model space is huge (2^p) we need to construct an efficient search procedure.

We generalize the Gaussian mixture prior approach in this non-Gaussian modeling framework. For non-Gaussian data it is well known that conjugate priors do not exist for regression coefficients. The computations are then potentially much harder, particularly when sampling from a large model space.

We exploit the use of a random residual component within the model. The use of a residual component is consistent with the belief that there may be unexplained sources of variation in the data, perhaps, due to explanatory variables that were not

recorded in the original study. By adopting a Gaussian residual effect many of the conditional distributions for the model parameters will be of standard form which greatly aids in computation.

In Chapter IV, we propose a new model–based approach for curve clustering. When gene expressions are observed by time, one main concern is now to cluster the gene expresson patterns over time. In this setting, clustering methods can be divided into two categories: non–model–based methods and model–based methods. In the non-model-based methods, such as hierarchical clustering (Eisen *et al.*,1998), clustering using correlation (Chu *et al.*,1999) and self-organising maps (Tamayo *et al.*,1999), the time dependency of the gene expression data is not considered, although time is a possible main factor in the level of gene expression. Model–based clustering methods are usually based on a finite mixture distribution, especially in the multivariate normal distribution (Yeung *et al,* 2001), where the number of clusters is determined by a model–choice criterion. A mixture–effects model with B–splines using an EM algorithm (Luan *et al.*, 2003) and hidden markov models (Schliep *et al.*, 2003) are examples of model–based methods.

We are motivated by Wakefield *et al.* (2003) who modelled the trajectory as a function of time and gene specific parameters, and clustered these curves using a reversible jump MCMC. Wakefield *et al.* (2003) used a first-order random walk model for gene-based parameters in a sporulation data (Chu *et al.*,1999) and a mixture of periodic function model for the cell–cycle data (Spellman *et al.*,1998)).

We propose a mixture of Dirichlet processes model using a discrete Wavelet transform for curve clustering as a fully Bayesian approach. Each iteration of the MCMC algorithm generates the cluster structure of these coefficients as a by-product (Escobar *et al.,*1998). We use a marginal posterior mode of their cluster memberships.

CHAPTER II

GENE SELECTION: A BAYESIAN VARIABLE SELECTION APPROACH*

2.1. Introduction

In this Chapter we will suggest a model based approach to this variable selection problem. Rather than fixing the dimension (the number of selected genes in the problem), we will assign it a prior distribution. This creates additional flexibility, as well as the ability to impose a constraint by limiting the dimension through the support of the prior. So, the prior works as a penalty to create this constraint. We will use a Markov Chain Monte Carlo (MCMC) (Gilks *et al.*, 1996) based stochastic search algorithm to identify the important genes. Though the model space is very large, as with p genes (say p=3000) we have 2^p models, and exhaustive computation over this model space is not possible. MCMC based stochastic search algorithms, which are less greedy and more efficient than most of the other alternatives, are successfully implemented to identify significant genes. We have considered the model for binary events only, but the extension to multi-category data is straightforward.

We will consider a data set from Hedenfalk *et al.* (2001) comparing the expression profiles of hereditary breast cancers. We want to identify the useful genes that can discriminate between BRCA1 and BRCA2 or sporadic breast cancers. The idea is to identify a small number of genes (by penalizing the dimension) that have the greatest discriminating power, thereby allowing researchers to quickly focus on the most promising candidates for diagnostics and therapeutics.

---

*Reprinted with permission from "Gene Selection: a Bayesian Variable Selection Approach" by Kyeong Eun Lee, Naijun Sha, Edward R. Dougherty, Marina Vannucci and Bani K. Mallick, 2003. *Bioinformatics*, Vol. 19, 90–97. 2003 by Oxford University Press.

2.2. Model for Gene Selection

We have binary responses,$\mathbf{Y}$, where Y_i=1 indicates the tumor sample i is class1 and $Y_i = 0$ indicates it is class2, for $i = 1, \cdots n$. For each sample, we have a measurement of the expression levels for all the genes, so X_{ij} is the measurement of the expression level of the jth gene for the ith sample where $j = 1, \cdots p$.

$$\begin{bmatrix} \text{Response} & \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene } p \\ Y_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Y_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

We assume that Y_i have the independent binary distributions so that $Y_i = 1$ with probability p_i, and it is independent of other $Y_j, j \neq i$. Then we relate the gene expression level with the response using a probit regression model which yields

$$Pr(Y_i = 1|\beta) = \Phi(X_i^{'}\beta)$$

where X_i is the ith row of the matrix X (vector of gene expression levels of the ith sample), β is the vector of regression parameters (β_j is the regression parameter corresponding to the jth gene) and Φ is the cumulative normal distribution function.

Albert and Chib (1994) introduced n independent latent variables $Z_1, \cdots, Z_n$ into the binary response regression, where $Z_i \sim N(X_i'\beta, 1)$

$$Y_i = \begin{cases} 1 & Z_i > 0 \\ 0 & Z_i \leq 0 \end{cases}$$

.

The latent variable has a linear model form as $Z_i = X_i'\beta + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$.

We now define γ to be the $p \times 1$ vector of indicator variables with ith element

γ_i such that $\gamma_i = 0$ if $\beta_i = 0$ (the gene is not selected) and $\gamma_i = 1$ if $\beta_i \neq 0$ (the gene is selected). Given γ, β_γ consists of all nonzero elements of β, and let X_γ be the columns of X corresponding to those elements of γ that are equal to 1. To complete the hierarchical model we need make the following prior assumptions:

(1) Given γ, the prior for $\beta_\gamma|\gamma$ is $\beta_\gamma \sim N(0, c(X'_\gamma X_\gamma)^{-1})$ where c is a positive scale factor specified by the user. After extensive testing, Smith and Kohn (1996) suggested to choose c between 10 to 100 for linear model problems. We will fix it to a large value, $c = 100$, so that the prior of β_γ, given γ, contains very little information about β_γ compared to the likelihood.

(2) The γ_i are assumed a *priori* to be a independent with $Pr(\gamma_i = 1) = \pi_i, 0 \leq \pi_i \leq 1$, for $i = 1, \cdots, p$. Now the value of π_i will be chosen to be small which will restrict the number of genes in the model. Here if we have prior knowledge that some genes are more important than others, we can implement computation easily by assigning larger or smaller value of π in the scale of importance 0 to 1. This prior is more useful when we will try to model a *priori* the interaction among the genes. The choice of a small value of π will indirectly restrain the number of selected genes in the model.

2.3. Computation

Due to the binary nature of the data we cannot obtain the posterior distribution in explicit form. So we use a MCMC method (Gilks *et al.*, 1996), specifically Gibbs sampling (Gelfand and Smith, 1990), to generate the posterior distributions of parameters.

Our unknowns are (Z, β, γ). To implement Gibbs sampling we need to simulate from the complete conditional distributions. Rather than drawing from the complete

conditional distributions we modify the algorithm and draw γ from the marginal distribution (integrating β) which speeds up the computation. It can be shown that this modified Gibbs sampler still leaves the target posterior distribution invariant. So our computation is:

(i) Draw from $\gamma|Z$, the marginalized conditional distribution obtained after integrating out β (this conditionally independent of Y). Now,

$$\begin{aligned} p(Z|\gamma) &\propto \int_{\beta} p(Z|\beta_\gamma)p(\beta_\gamma|\gamma)d\beta_\gamma \\ &\propto \exp[-1/2(Z'Z - \frac{c}{1+c}Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z)]. \end{aligned}$$

The proof is a simple application of Bayesian linear model theory (Lindley and Smith, 1972; Denison *et al.*, 2002) and provided in the appendix.

The conditional distribution of $\gamma|Z$ is

$$\begin{aligned} p(\gamma|Z) &\propto p(Z|\gamma)p(\gamma) \\ &\propto \exp[-1/2(Z'Z - \frac{c}{1+c}Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z)]\Pi_{i=1}^{p}{\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i} \end{aligned}$$

Rather than sampling γ as a vector, it is better to draw component wise of the vector, γ, from $p(\gamma_i|Z, \gamma_{j\neq i})$ which is

$$\begin{aligned} p(\gamma_i|Z, \gamma_{j\neq i}) &\propto p(Z|\gamma)p(\gamma_i) \\ &\propto \exp[-1/2(Z'Z - \frac{c}{1+c}Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z)]{\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \end{aligned}$$

(ii) Draw from $\beta_\gamma|\gamma, Z$, (which is conditionally independent of Y), $p(\beta|\gamma, Z) \sim N(V_\gamma X'_\gamma Z_\gamma, V_\gamma)$ where $V_\gamma = \frac{c}{1+c}(X'_\gamma X_\gamma)^{-1}$ and the index vector γ means all the ele-

ments only corresponding to $\gamma_i = 1, i = 1, \cdots, n$.

(iii) The full conditional distribution of Z_i is as follows:

$$\begin{aligned} Z_i|\beta, Y_i = 1 &\propto N(X_i'\beta, 1) \qquad \text{truncated at the left by 0} \\ Z_i|\beta, Y_i = 0 &\propto N(X_i'\beta, 1) \qquad \text{truncated at the right by 0} \end{aligned}$$

The distribution of Z is a truncated normal and can be generated using Robert's (Robert, 1995) optimal exponential accept-reject algorithm.

After suitable burn-in period (usually 10,000) we obtained the MCMC samples at the t-th iteration as $\{\beta^{(t)}, Z^{(t)}, \gamma^{(t)},\ t = 1, \cdots, m\}$. We can use these samples from the posterior distributions for posterior inference and prediction.

The Algorithm

Start with initial values $[\gamma^{(0)}, Z^{(0)}, \beta^{(0)}]$

At the tth iteration

(i) Draw $\gamma^{(t)}$ from $p(\gamma|Z^{(t-1)})$.

(ii) Draw $Z^{(t)}$ from $p(Z|\beta^{(t-1)}, \gamma^{(t)})$.

(iii) Draw $\beta^{(t)}$ from $p(\beta|Z^{(t)}, \gamma^{(t)})$.

(iv) Let $t \leftarrow t + 1$.

Continue required number of iterations.

Stop

For decision making we can calculate the relative number of times each gene appeared in the MCMC sample (number of time the corresponding γ is 1). This will give us an estimate of the posterior probability of inclusion of that gene and tell us the relative importance of the gene for classification purposes.

We can also obtain the predictive classification of a new observation Y_{new} condi-

tion on the expression levels as

$$P(Y_{new} = 1|X) = \int_Z \int_\beta p(Y_{new} = 1|X, Z, \beta) p(z, \beta|Y) dZ d\beta \tag{2.1}$$

and the Monte-Carlo estimate of this probability will be

$$\hat{P}(Y_{new} = 1|X) = \frac{1}{m} \sum_{t=1}^{m} p(Y_{new} = 1|X, Z^{(t)}, \beta^{(t)}, \gamma^{(t)})$$

and can be easily evaluated using normal cumulative distribution function.

2.4. Application of Gene Selection to Hereditary Breast Cancer Data

We apply the proposed strategy for discovering significant genes to a published data set (Hedenfalk *et al.*,2001) consisting of patients carrying mutations in predispoing genes, BRCA1, and patients not expected to carry a hereditary predisposing mutation. Pathological and genetic differences appear to imply different but overlapping functions for BRCA1 and BRCA2. In Hedenfalk *et al.* (2001), cDNA microarrays were used in conjuction with classification algorithms to show the feasiblity of using differences in global gene expression profiles to separate BRCA1 and BRCA2 type. They examined 22 breast tumor samples from 21 breast cancer patients. All patients except one were women; fifteen of the women had hereditary breast cancer, seven tumors with BRCA1 and eight tumors with BRCA2. In the analysis, 3226 genes were used for each breast tumor sample. We used our method to classify BRCA1 versus non–BRCA1 (BRCA2 and sporadic) samples.

We used two sample t–statistics to identify the starting value for γ by identifying, say, the five most significant genes. We then ran the MCMC sampler, in particular, the Gibbs sampling approach fixing $\pi_i = .005$ for all $i = 1 \cdots p$. The chain moved quite frequently and we used 50,000 iterations after a 10,000 iteration burn in period.

In Table I, we present the strongly significant genes with the largest frequencies.

We note that the three leading genes in Table I appear among the six strongest genes in an analogous list in Kim *et al.* (2002). This has occurred even though the rating in the latter paper is based upon the ability of a gene to contribute to a linear classifier, which is quite different than the criterion here. One of the strongest selected genes is keratin 8 (KRT8) is a member of the cytokeratin family of genes. Cytokeratins are frequently used to identify breast cancer metastases by immunohistochemistry, and keratin 8 abundance has been shown to correlate well with node-positive disease (Brotherick *et al.*, 1998). The gene TOB1 is the first in Table I, and appeared fifth in Kim *et al.* (2002). It interacts with the oncogene receptor ERBB2, and is found to be more highly expressed in BRCA2 and sporadic cancers, which are likewise more likely to harbor ERBB2 (Matsuda *et al.*, 1996). We note that the gene for the receptor was not on the arrays, so that the gene-selection algorithm was blinded to its input. Lastly, the second gene in Table I, appears as the sixth gene in the list of Kim *et al.* (2002).

Heat maps have become popular in the microarray literature, as graphical representations of the primary data where each point is associated with a color that reflects its value. Increasingly positive values are represented with reds of increasing intensity, overexpressed, and increasing negative values with greens of increasing intensity, underexpressed. A heat map of the identified genes is Figure 2. From this Figure 2, we are able to detect some patterns which indicates existence of different classes.

Table I. Strongly Significant Genes Found for Classifying BRCA1 versus Non BRCA1

Freq.	Image Clone ID	Gene Description
8.4	823940	"transducer of ERBB2, 1"
7.8	26184	"phosphofructokinase, platelet"
7.5	840702	SELENOPHOSPHATE SYNTHETASE ; Human selenium donor protein
7.1	897781	keratin 8
7.1	376516	cell division cycle 4-like
6.9	47542	small nuclear ribonucleoprotein D1 polypeptide (16kD)
6.6	366647	butyrate response factor 1 (EGF-response factor 1)
6.6	293104	phytanoyl-CoA hydroxylase (Refsum disease)
6.2	28012	O-linked N-acetylglucosamine (GlcNAc) transferase
6.1	212198	"tumor protein p53-binding protein, 2"
5.9	247818	ESTs
5.5	26082	very low density lipoprotein receptor
5.4	667598	PC4 and SFRS1 interacting protein 1
5.2	30093	RAN binding protein 1
5.1	73531	nitrogen fixation cluster-like
5	950682	"phosphofructokinase, platelet"
5	47681	"splicing factor, arginine/serine-rich (transformer 2 Drosophila homolog)"
4.9	46019	minichromosome maintenance deficient (S. cerevisiae) 7
4.9	307843	ESTs
4.8	548957	"general transcription factor II, i, pseudogene 1"
4.7	788721	KIAA0090 protein
4.7	843076	signal transducing adaptor molecule (SH3 domain and ITAM motif)
4.7	204897	"phospholipase C, gamma 2 (phosphatidylinositol-specific)"
4.7	812227	"solute carrier family 9, isoform 1"
4.6	566887	heterochromatin-like protein 1
4.6	563598	"gamma-aminobutyric acid (GABA) A receptor, pi"
4.5	324210	sigma receptor (SR31747 binding protein 1)

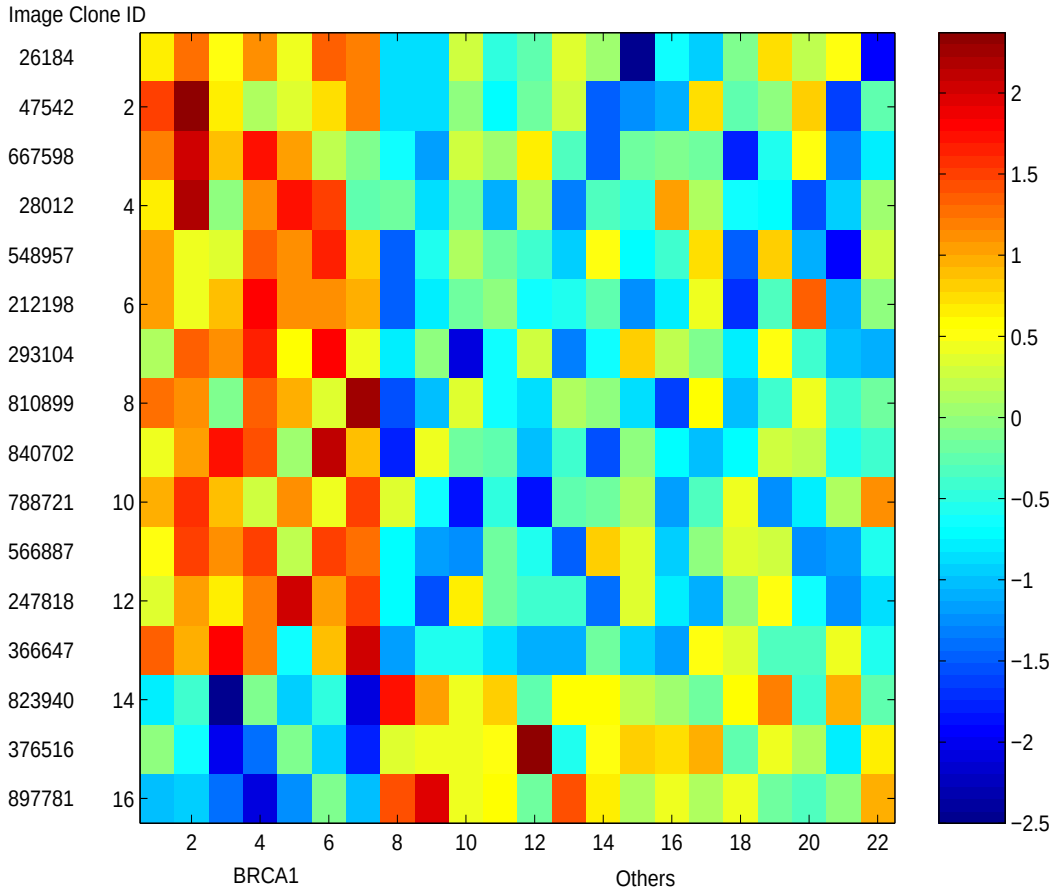

Fig. 2. Breat Cancer Data: Heat Map

We can check the model adequacy in two ways. (i) Cross validation: we use the cross-validation predictive density (Wilks *et al.*, 1996),

$$\begin{aligned}\hat{f}(y_i|Y_{-i}) &= \frac{1}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{f(y_i|\mathbf{Y}_{-i},\boldsymbol{\theta}^*_m)}} \\ &= \frac{1}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{f(y_i|\boldsymbol{\theta}^*_m)}},\end{aligned}$$

where $\boldsymbol{\theta}^*_m$ is the mth MCMC sample, and predict $Y = 1$ for that point if $\hat{f}(y_i = 1|Y_{-i}) > .5$. The second equality holds due to the conditional independency of y_i given $\boldsymbol{\theta}$.

(ii) Deviance: Deviance calculation is one of many criterion based methods that measures goodness of fit (McCullagh and Nelder, 1983). Lower deviance means better fit. We have calculated the probabilities and the deviance measures for different models in Table II which shows the adequacy of the models.

Model1: Using all genes in Table I

Model2: Using genes with frequencies more than 5%

Model3 : Using genes with frequencies more than 6%

Model4 : Using genes with frequencies more than 7%

We compared our cross validation results with other popular classification algorithms including feed forward neural networks, k-nearest neighbors, support vector machines (SVM) result in Table III. All other methods in Table III have used 51 genes (which we think is too many with respect to a sample size of 22) which may

produce instability in the classification process. Our procedure has used far fewer genes though the results are competitive to these other methods.

All the analyses we have presented are mainly marginal analyses of genes. The best model (with respect to minimum deviance criterion) we picked from all the samples is a 4 variable model and presented in Table IV.

Table II. Cross Validation of Breast Cancer Data

Y	Model 1 Pr(Y=1)	Model 2 Pr(Y=1)	Model 3 Pr(Y=1)	Model 4 Pr(Y=1)
1	1	1	0.9993	0.9998
1	1	1	1	0.9969
1	1	1	0.9999	1
1	1	1	0.9999	0.8605
1	1	1	0.9999	0.7766
1	1	1	0.9998	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0.0002
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0.0002
0	0	0	0.0018	0.0867
0	0	0	0.0005	0.007
0	0	0	0	0
0	0	0	0	0.2864
1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
Deviance	$1.2683e-12$	$3.1464e-7$	0.0071	1.6843
Number of misclassifications	0	0	0	1

Table III. Cross Validation Errors of Different Models for Breast Cancer Data

	Model	Cross-Validation Error*
1	Feed-forward Neural Networks	0
	(3 hidden neurons, 1 hidden layer)	(An average of 1.5 misclassifications)
2	Gaussian Kernel	1
3	Epanechnikov Kernel	1
4	Moving Window Kernel	2
5	Probabilistic Neural Network (r=0.01)	3
6	kNN(k=1)	4
7	SVM Linear	4
8	Perceptron	5
9	SVM Nonlinear	6

$*$: Number of Misclassified Samples Feature Selection: 51 Features used in the paper 'Gene-Expression Profiles in hereditary Breast Cancer', Vol 344, NJE

2.4.1. Sensitivity Analysis

We have checked the sensitivity of our analysis to the measurement of expression levels (as they are always subject to measurement errors) by adding Gaussian noises to the expression values. We reanalyzed the data contaminated by different levels of Gaussian noise to obtain the newly selected genes and have reproduced them in Table V.

It is clear from the table that our analysis is stable as it is selecting almost similar genes under different noise levels over the expression values. Among the seven leading genes in Table I , the following appear across all noise conditions in Table V : Keratin

Table IV. The Best Model in the Breast Cancer Data

Image Clone ID	Gene Description
HV5H10	"transducer of ERBB2, 1"
HV16C12	pyruvate dehydrogenase (lipoamide) beta
HV33F6	"protein phosphatase 3 (formerly 2B), regulatory subunit B"
HV44D9	hypothetical protein PRO0823
Deviance	0.0071

8, phosphofructokinase platelet, Selenophosphate synthetase, and butyrate response factor 1. TOB1 is only omitted at the highest noise level.

To check the prior sensitivity we have rerun our algorithm for several choices of c between 10 and 100 and the results are not that sensitive towards the choice of c. We suggest to fix a large value of c (say 100) as it is almost a non-informative prior. The number of genes selected are very sensitive towards the choice of π. On average the number of genes selected will be $m \times \pi$ where m is the total number of genes. For our case $m = 3226$ and sample size is only 23. With this small sample size we don't want to select too many genes (not more than 23) and we can restrain the number of selected genes by choosing π to be small. For example if we want to keep the number of selected genes around 23 the choice of π should be 0.007. This way the Bayesian method is successful to penalize the number of selected genes through the help of the prior specifications. We reanalyzed the data for several choices of π from 0.001 to 0.1 which selects different number of genes in different cases but the identification of the frequency arising genes remained same.

Table V. Breast Cancer Data: Sensitivity Analysis

Error $\sim N(0, 0.1^2)$		Error $\sim N(0, 0.2^2)$		Error $\sim N(0, 0.5^2)$	
Freq.	Image Clone ID	Freq.	Image Clone ID	Freq.	Image Clone ID
10.5	840702*	10.5	26184*	12.7	840702*
9.5	897781*	10.1	840702*	10.4	26184*
8.6	247818*	9.6	897781*	9.0	293104*
8.3	26184*	7.6	566887	9.0	897781*
7.7	212198*	7.6	293104*	8.0	247818*
7.5	307843*	7.3	46019*	7.8	307843*
7.4	47681*	7.3	212198*	7.8	566887*
6.8	293104*	6.9	247818*	7.4	548957*
6.3	823940*	6.8	564803	7.1	46019*
5.7	566887*	6.3	788721*	7.1	810899
5.7	28012*	6.0	366647*	6.6	46182
5.6	376516*	5.9	307843*	6.5	47681*
5.5	46019*	5.9	73531*	6.4	366647*
5.4	548957*	5.8	825478*	6.4	28012*
5.3	26082*	5.8	28012*	6.3	843076*
5.3	46182	5.4	376516*	6.0	26082*
5.3	30093*	5.3	204897*	5.9	788721*
5.1	366647*	5.2	26082*	5.8	667598*
5.0	248531	5.2	248531	5.7	212198*
4.9	246524	5.2	47681*	5.6	73531*
4.9	204897*	5.1	667598*	5.5	30093*
4.7	139540	5.0	810899*	5.3	825478*
4.7	47542*	5.0	823940*	5.3	246524
4.4	32790	4.8	843076*	5.1	564803
4.4	134748	4.8	46182	5.0	248531
4.3	810899*	4.7	246524	4.9	897646
4.2	667598*	4.7	324210*	4.8	950682

$*$: Selected genes which were already in the original analysis.

2.5. Application to Leukemia Data

Now we apply our method to a larger data set with a test set where we can perform our prediction validation. The leukemia data set was described by Golub *et al.* (1999). Bone marrow or peripheral blood samples were taken from 72 patients with either myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). Following the experimental setup of the original paper, the data were split into a training set consisting of 38 samples of which 27 are ALL and 11 are AML, and a test set of 34 samples, 20 ALL and 14 AML. The dataset contains expression levels for 7129 human genes produced by Affymetrix high-density oligonucleotide microarrays. The scores in the dataset represent the intensity of gene expression after being rescaled to make overall intensities. Golub *et al.* (1999) used a predictor trained using a weighted voting scheme on the training samples, and classifyed correctly on all samples for which a prediction is made, 29 of the 34, declining to predict for five samples. We performed our anlaysis with the same choices of the hyper parameters as in the first example and report here our resluts. In Table VI, we provide the genes which appeared more frequently in our posterior samples. There are several genes, including the top one, that also belong to the set of 50 genes used by Golub *et al.* (these genes are reported with asterisks in Table VI). We considered five models based on marginal frequencies as follows:

Model1 : Using genes with frequencies more than 115

Model2 : Using genes with frequencies more than 125

Model3 : Using genes with frequencies more than 150

Model4 : Using genes with frequencies more than 200

Model5 : Using genes with frequencies more than 250

We have calculated the probabilities and the deviance measures for different models in Table VII which shows the adequacy of the models. We used the genes that appeared more than 250 (appeared to be the top 5 genes) to perform predictions on the test data. The prediction results are reported in Table VIII. Only one observation is misclassified (the observation number 29). These results appear to improve predictions made by Golub *et al.* (1999) and use only 5 genes (Model5) rather than 50. Figure 3 shows the heat map of the 27 genes (Model1) identified by our methods which differentiate two different classes, ALL and AML.

2.6. Discussion

We have proposed a Bayesian model to identify important genes using expression level data, by forming the problem as a variable selection problem for binary data. We have used a hierarchical probit model and MCMC based stochastic search techniques to obtain the posterior distributions. We have proposed one based on Gibbs sampling. Though ideally the Metropolis–Hastings algorithm should be faster, it has a tendency to stick a region due to the high dimensionality of the problem. We have mainly used the Gibbs algorithm in this paper but in future we will investigate the more adaptive MH algorithm (or a mixture of two) to speed up our computation.

Here we have fixed the π value but we can extend our model assuming π is an unknown model parameter. Assigning a conjugate beta distribution prior on π, the extension is straightforward.

We have assumed the genes are independent but in our framework we can very easily extend it for dependent cases. For example: consider if the event where the ith gene expression increases the chance that jth gene will be expressed. In our framework we can account for the dependence through the prior distribution of γ.

Rather than assuming the γ_i are independently distributed we can use a Markov model whose transition matrices will be defined as $p(\gamma_j = 1|\gamma_i = 1)$. This type of problem will be handled in future research.

In this paper we have considered binary data. Extension to more than two categories can be found in Albert and Chib (1993) and development of a variable selection model in that setup is in Sha (2002).

Table VI. Leukemia Data: Strongly Significant Genes Found for Classification

Freq.	ID	Gene Description
77.2	1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)*
48.5	760	CYSTATIN A
28.3	2288	DF D component of complement (adipsin)*
27.8	4847	Zyxin*
26.8	1144	"SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)"
24.2	1120	SNRPN Small nuclear ribonucleoprotein polypeptide N
21.1	4535	RETINOBLASTOMA BINDING PROTEIN P48
19.8	6218	"ELA2 Elastatse 2, neutrophil"
19.5	6200	Interleukin 8 (IL8) gene *
19.5	1834	CD33 CD33 antigen (differentiation antigen)*
18.8	1630	Inducible protein mRNA*
17.9	5772	C-myb gene extracted from Human (c-myb) gene*
16.9	1745	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog*
16.7	804	Macmarcks
16.1	2354	CCND3 Cyclin D3 *
14.9	3252	"GLUTATHIONE S-TRANSFERASE, MICROSOMAL"
13.5	6201	INTERLEUKIN-8 PRECURSOR*
13.5	1685	Terminal transferase mRNA
13.1	6041	APLP2 Amyloid beta (A4) precursor-like protein 2
13.1	1779	MPO Myeloperoxidase
12.7	6855	TCF3 Transcription factor 3
12.6	173	"PRKCD Protein kinase C, delta"
12	2642	MB-1 gene*
12	1829	PPGB Protective protein for beta-galactosidase
11.9	4107	PLECKSTRIN
11.8	697	"KIAA0235 gene, partial cds"
11.7	229	KIAA0102 gene

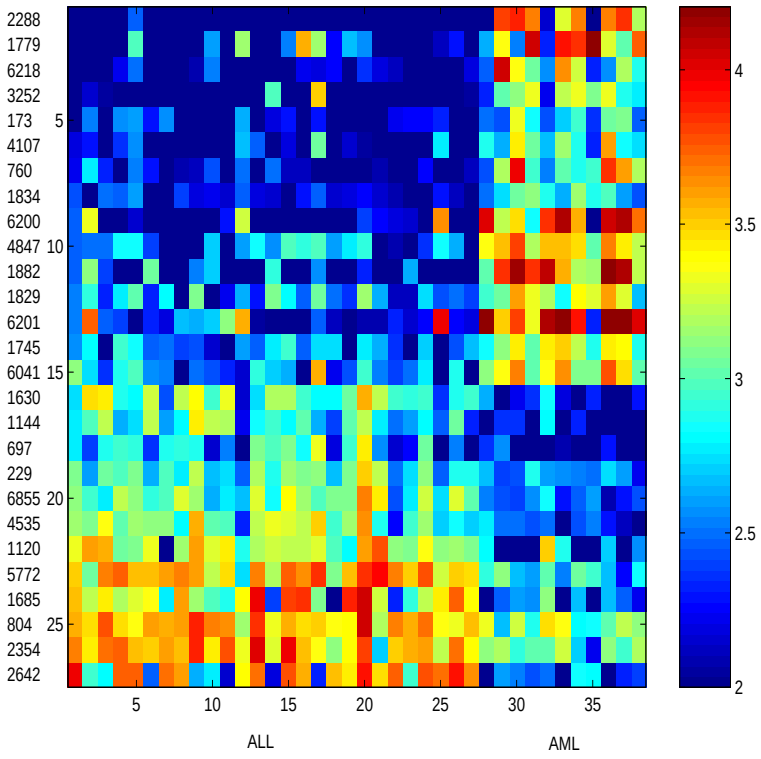

Fig. 3. Leukemia Data: Heat Map

Table VII. Crossvalidated Classification Probabilities and Deviance for the Leukemia Data

Y	$\Phi(X_1\hat{\beta}_1)$	$\Phi(X_2\hat{\beta}_2)$	$\Phi(X_3\hat{\beta}_3)$	$\Phi(X_4\hat{\beta}_4)$	$\Phi(X_5\hat{\beta}_5)$
1	1	1	1	1	1
1	1	0.9998	0.9993	0.9982	0.9813
1	1	1	1	1	1
1	1	1	1	0.9998	1
1	1	1	0.9998	0.9985	0.9938
1	1	1	0.9999	0.9996	1
1	1	0.9999	0.9996	1	1
1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	1	0.9958	0.9933
1	1	1	1	1	1
1	1	1	0.9991	0.9993	0.9999
1	1	1	1	0.9999	0.9999
1	1	1	0.9988	0.994	0.9974
1	1	1	1	0.9994	0.9969
1	1	1	1	0.9997	0.9985
1	1	0.9999	0.9959	0.9909	0.9879
1	1	1	0.9999	1	1
1	1	1	1	0.9987	1
1	1	1	1	1	0.9986
1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	0.9994	1	1
1	1	1	1	1	1
1	1	1	1	0.9999	1
1	1	1	1	1	1
1	1	1	1	1	1
0	0	0	0.0037	0.0158	0.0083
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0.0011
0	0	0.0001	0.0005	0.0004	0
0	0	0	0	0	0
0	0	0	0	0	0.0001
0	0	0.0001	0	0	0.0174
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0.0004	0.0003
Deviance	3.53E-08	0.0014	0.025	0.0863	0.1605

Table VIII. Leukemia Data: Prediction on the Test Set Using Genes with Frequencies Higher than 250

Y	$\hat{\Phi}(X_{test}\hat{\beta})$	Y	$\hat{\Phi}(X_{test}\hat{\beta})$
1	1.0000	1	0.2503
1	1.0000	1	1.0000
1	1.0000	1	1.0000
1	0.9972	1	0.9999
1	1.0000	1	1.0000
1	1.0000	1	1.0000
1	1.0000		
1	1.0000		
1	1.0000		
1	1.0000		
1	1.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
1	0.9963		
1	1.0000		
0	0.0000		
0	0.0000		
1	1.0000		
0	0.0000		
0	0.1143		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0612		

CHAPTER III

BAYESIAN METHODS FOR VARIABLE SELECTION IN THE SURVIVAL MODEL FOR APPLICATION TO DNA MICROARRAY DATA

3.1. Introduction

In this chapter, we suggest a variable selection technique for survival models using a Bayesian model–based variable selection approach. Typical Bayesian variable selection methods are based on the assumptions of Gaussian distributions for the likelihood and use of conjugate mixture priors to obtain marginal distributions (George and McCulloch, 1993). We will extend such models to the survival data context where the responses are time to event. We will address the issue of how to select the significant predictors as well as assess the survival curves using the Weibull regression or proportional hazards model (PH) where the sample size is much smaller than the number of predictors (genes).

Current techniques used for variable selection in survival models include asymptotic procedures based on score tests and other approximate chi-square procedures. These asymptotic methods may not be valid here because n is much smaller than p. The literature on Bayesian variable selection for survival models is still sparse. Raftery, Madigan and Volinsky (1995) had used the approximate BIC criterion for variable selection. Ibrahim and Chen (2000) advocate a predictive approach which is efficient with a small number of predictors (p is small). As our model space is large (2^p), we need to construct an efficient selection procedure.

We will generalize the Gaussian mixture prior approach to this non-Gaussian modeling framework. For non-Gaussian data it is well known that conjugate priors do not exist for the regression coefficients. Computation is then potentially much harder

particularly when sampling the dimension of the model space. This is due to possibly strong posterior correlation between the elements of regression parameters such that adding or removing a variable can result in a large drop in the model likelihood unless careful update proposals are made to the coefficients to accommodate the change. The construction of good proposals is not trivial and depends on both the form of the model and on the data.

In this chapter we exploit the use of a random residual component within the model. The use of a residual component is consistent with the belief that there may be unexplained sources of variation in the data, perhaps due to explanatory variables, that were not recorded in the original study. By adopting a Gaussian residual effect, many of the conditional distributions for the model parameters will be of standard form which greatly aids in the computations.

It is quite a new approach to do a variable selection in the survival model when the sample size is much smaller than the number of variables. Our proposed method can be particularly useful for the general large p small n situations and not just for DNA microarray data.

We will consider two cDNA data sets: B-cell lymphoma data set(Alizadeh *et al.* 2000) and breast carcinoma samples (Sørlie *et al.* 2001). We want to identify a set of responsible genes which explain the survival time in each data set.

3.2. Weibull Regression Model

Let T_i be the survival time (observed or censored) for the ith patient and X_{ij}s are the $p+1$ covariates associated with it. Usually X_{i0} indicates the binary or multi-category phenotype covariate and other X_{ij}s are p gene expressions from DNA microarray data, which is continuous in nature.

$$\begin{bmatrix} \text{Survival Time} \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} \begin{bmatrix} \text{Category} & \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene p} \\ X_{10} & X_{11} & X_{12} & \cdots & X_{1p} \\ X_{20} & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

In this section, we assume a parametric model, say, the Weibull distribution for the survival time. Suppose we have independently distributed survival time, $\boldsymbol{t} = (t_1, \cdots, t_n)'$, which follows the Weibull distribution with a shape parameter α and a scale parameter γ_i^*.

$$f(t|\alpha, \gamma^*) = \begin{cases} \alpha\gamma^* t^{\alpha-1} \exp(-\gamma^* t^\alpha) & \text{for } t > 0, \alpha > 0, \gamma^* > 0 \\ 0 & \text{o.w.} \end{cases}$$

It is more convenient to write the model in terms of the reparametrization $\lambda = \log(\gamma^*)$, and its corresponding pdf is

$$f(t|\alpha, \lambda) = \alpha t^{\alpha-1} \exp(\lambda - \exp(\lambda) t^\alpha) I(t > 0, \alpha > 0),$$

its hazard function is $h(t|\alpha, \lambda) = \alpha \exp(\lambda) t^{\alpha-1}$ and its survival function is $S(t|\alpha, \lambda) = \exp(-\exp(\lambda) t^\alpha)$. The censoring indicator variables are $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)'$, where

$$\delta_i = \begin{cases} 0 & \text{if } t_i \text{ is right censored} \\ 1 & \text{if } t_i \text{ is a death time} \end{cases}.$$

The likelihood function of λ and α :

$$\begin{aligned} L(\alpha, \boldsymbol{\lambda}|\mathbf{D}) &= \prod_{i=1}^{n} f(y_i|\alpha, \lambda_i)^{\delta_i} S(y_i|\alpha, \lambda_i)^{(1-\delta_i)} \\ &= \alpha^d \exp\left\{ \sum_{i=1}^{n} (\delta_i \lambda_i + \delta_i(\alpha_i) \log(y_i)) - \sum_{i=1}^{n} (\exp(\lambda_i) y_i^\alpha) \right\} \end{aligned} \tag{3.1}$$

where $\mathbf{D} = (n, \boldsymbol{t}, \boldsymbol{\delta})$.

In order to build up a Weibull regression model and perform the variable selec-

tion, we developed a hierarchical model introducing the covariate $\boldsymbol{x}$ (gene expression) through $\boldsymbol{\lambda}$ as $\lambda_i = \mathbf{X}_i'\boldsymbol{\beta}$. As mentioned earlier, with a complicated likelihood as in (1), it is hard to identify a conjugate mixture prior distribution to perform the variable selection. We overcome the computational difficulty by including a random residual component ϵ_i. The model is modified as

$$\lambda_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2).$$

This introduction of ϵ enables to generate samples from full conditionals of all other parameters which is consistent with the belief that there may be unexplained sources of variation in the data, perhaps departure from the assumption of linearity. Now we can construct the Gaussian mixture prior for $\boldsymbol{\beta}$ to perform the variable selection procedure. Define $\boldsymbol{\gamma}$ to be an arbitrary $p \times 1$ vector of indicator variables with ith element γ_i such that $\gamma_i = 0$ if $\beta_i = 0$ (the gene is not selected) and $\gamma_i = 1$ if $\beta_i \neq 0$ (the gene is selected). Given γ, let β_γ consist of all nonzero elements of β and let $\boldsymbol{x}_\gamma$ be the columns of $\boldsymbol{x}$ corresponding to those elements of γ that are equal to 1. To complete the hierarchical model we need make the prior assumptions:

1. Given γ, the prior for β_γ will be $\beta_\gamma \sim \mathrm{N}\{0, c(\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\}$, where c is a positive scale factor specified by the user. Smith and Kohn (1996) suggested to choose c between 10 to 100 for linear model problems. We will fix $c = 100$, so that the prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, given $\boldsymbol{\gamma}$, contains little information about $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

2. The γ_i will be assumed to be a priori independent with $\mathrm{pr}(\gamma_i = 1) = \pi_i$. The values of π_i will be chosen to be small which will restrict the number of genes in the model. For example if we have 3000 total number of genes and want to allow only 15 genes due to small sample size, then we will fix $\pi_i \equiv 0.005$ to achieve the purpose. In addition, if we have prior knowledge that some genes

are more important than others, we can incorporate this easily by assigning larger values of π.

So our hierarchical structure of the model for variable selection is as follows:

$$
\begin{aligned}
[T_i|\alpha, \lambda_i] &\sim \text{Weibull}(\alpha, \lambda_i) \\
[\lambda_i|\beta, \sigma] &\sim \text{Gaussian}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2) \\
[\beta_i|\gamma_i] &\sim \text{Gaussian}(0, \gamma_i c\sigma^2) \\
[\gamma_i] &\sim \text{Bernoulli}(\pi_i) \\
[\alpha] &\sim \text{Gamma}(\alpha_0, \kappa_0) \\
[\sigma^2] &\sim \text{InverseGamma}\left(a_0, \frac{b_0}{2}\right).
\end{aligned}
$$

As the posterior distributions of the parameters are not of explicit form, we need to use Markov Chain Monte Carlo (MCMC) based approaches, specifically Gibbs sampling and Metropolis algorithms to generate samples from the posterior distribution. All the conditional distributions required for MCMC are obtained in the next section.

3.2.1. Conditional Distributions and Posterior Sampling

To perform Gibbs sampling we need to obtain the conditional distributions of the parameters as follows.

For the convenience, let $\theta = \log(\alpha)$. The conditional distribution of λ is

$$
\begin{aligned}
p(\lambda|\mathbf{D}, \beta_\gamma, \theta, \sigma^2) &\propto \exp\left\{\theta d + \sum_{i=1}^{n}(\nu_i\lambda_i + \nu_i(\mathrm{e}^\theta - 1)\log(y_i)) - \sum_{i=1}^{n}(\mathrm{e}^{\lambda_i} y_i^{\mathrm{e}^\theta})\right\} \\
&\times \exp\left\{-\frac{1}{2\sigma^2}(\lambda - X\gamma\beta_\gamma)'(\lambda - X\gamma\beta_\gamma)\right\}
\end{aligned}
$$

Since λ_i's are conditionally independent, it is convenient to draw componentwise.

$$\begin{aligned} p(\lambda_i|\mathbf{D}, \boldsymbol{\lambda}_{j\neq i}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \theta, \sigma^2) &\propto \exp\left\{\theta d + \nu_i\lambda_i + \nu_i(\mathrm{e}^\theta - 1)\log(y_i) - \exp(\lambda_i)y_i^{\mathrm{e}^\theta}\right\} \\ &\times \frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\lambda_i - \mathbf{X}'_{\boldsymbol{\gamma},i}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^2\right\} \end{aligned}$$

where $\mathbf{X}_{\boldsymbol{\gamma},i}$ is the ith column of $\mathbf{X}_{\boldsymbol{\gamma}}$.

The conditional distribution of θ is

$$\begin{aligned} p(\theta|\mathbf{D}, \boldsymbol{\lambda}, \beta_\gamma, \sigma^2) &\propto \exp\left\{\theta d + \sum_{i=1}^{n}(\nu_i\lambda_i + \nu_i(\mathrm{e}^\theta - 1)\log(y_i)) - \sum_{i=1}^{n}(\mathrm{e}^{\lambda_i}y_i^{\mathrm{e}^\theta})\right\} \\ &\times \mathrm{e}^{\theta(\alpha_0-1)}\exp(-\kappa_0\mathrm{e}^\theta)\mathrm{e}^\theta \end{aligned}$$

In order to speed up the computation, we draw $\boldsymbol{\gamma}$ from the marginal distribution(integrating $\boldsymbol{\beta}$ out). Since $p(\boldsymbol{\lambda}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2) = p(\boldsymbol{\lambda}|\beta_\gamma, \sigma^2)p(\beta_\gamma|\sigma^2)$,

$$\begin{aligned} p(\boldsymbol{\lambda}|\beta_\gamma, \sigma^2)p(\beta_\gamma|\sigma^2) &\propto \frac{|X'_\gamma X_\gamma|^{-1/2}}{(c\sigma^2)^{q_\gamma/2}}\frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(1+c^{-1})\beta'_\gamma(X'_\gamma X_\gamma)\beta_\gamma + \lambda' X_\gamma\beta_\gamma\right\} \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\lambda'\lambda\right) \\ &= \exp\left(-\frac{1}{2}\beta'_\gamma V_\gamma^{-1}\beta_\gamma + \beta'_0 V_\gamma^{-1}\beta_\gamma - \frac{1}{2}\beta'_0 V_\gamma^{-1}\beta_0\right) \\ &\quad \times \frac{|X'_\gamma X_\gamma|^{-1/2}}{(c\sigma^2)^{q_\gamma/2}}\exp\left(-\frac{1}{2\sigma^2}\lambda'\lambda\right) \end{aligned}$$

where $\mathbf{V}_{\boldsymbol{\gamma}} = \sigma^2\frac{c}{1+c}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ and $\boldsymbol{\beta}_0 = \frac{c}{1+c}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}'_{\boldsymbol{\gamma}}\boldsymbol{\lambda}$.

So

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{\gamma}, \sigma^2) &\propto \int_{\boldsymbol{\beta}} p(\boldsymbol{\lambda}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \sigma^2)d\boldsymbol{\beta}_{\boldsymbol{\gamma}} \\ &\propto \left(\frac{1}{1+c}\right)^{q_{\boldsymbol{\gamma}}/2}\exp\left\{-\frac{1}{2\sigma^2}S(\boldsymbol{\gamma})\right\} \end{aligned}$$

where $S(\boldsymbol{\gamma}) = \boldsymbol{\lambda}'\boldsymbol{\lambda} - \frac{c}{1+c}\boldsymbol{\lambda}'\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}'_{\boldsymbol{\gamma}}\boldsymbol{\lambda}$.

Therefore, the marginal distribution of $\boldsymbol{\gamma}$ is

$$p(\gamma|\lambda, \sigma^2) \propto p(\lambda|\gamma, \sigma^2)p(\gamma)$$

$$\propto \quad (1+c)^{-q\boldsymbol{\gamma}/2}\exp\left\{-\frac{1}{2\sigma^2}S(\gamma)\right\}\prod_{i=1}^{n}{\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i}.$$

Now rather than drawing γ as a vector, it is better to draw component wise from $p(\gamma_i|\lambda,\gamma_{j\neq i})$ which is

$$\begin{aligned} p(\gamma_i|\boldsymbol{\lambda},\boldsymbol{\gamma}_{j\neq i},\sigma^2) &\propto p(\boldsymbol{\lambda}|\boldsymbol{\gamma},\sigma^2)p(\gamma_i) \\ &\propto (1+c)^{-q\boldsymbol{\gamma}/2}\exp\left\{-\frac{1}{2\sigma^2}S(\boldsymbol{\gamma})\right\}{\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \end{aligned}$$

Since

$$\begin{aligned} p(\gamma_i=1|\boldsymbol{\lambda},\boldsymbol{\gamma}_{j\neq i},\sigma^2) &\propto \pi_i(1+c)^{-q{\boldsymbol{\gamma}_i}^1/2}\exp\left\{-\frac{1}{2\sigma^2}S({\boldsymbol{\gamma}_i}^1)\right\} \\ p(\gamma_i=0|\boldsymbol{\lambda},\boldsymbol{\gamma}_{j\neq i},\sigma^2) &\propto (1-\pi_i)(1+c)^{-q{\boldsymbol{\gamma}_i}^0/2}\exp\left\{-\frac{1}{2\sigma^2}S({\boldsymbol{\gamma}_i}^0))\right\} \end{aligned}$$

where ${\boldsymbol{\gamma}_i}^1=(\gamma_1,\cdots,\gamma_i=1,\cdots,\gamma_p)$ and ${\boldsymbol{\gamma}_i}^0=(\gamma_1,\cdots,\gamma_i=0,\cdots,\gamma_p)$,

$$p(\gamma_i=1|\boldsymbol{\lambda},\gamma_{j\neq i},\sigma^2)=\frac{1}{1+\dfrac{1-\pi_i}{\pi_i}(1+c)^{1/2}\exp\{-\dfrac{1}{2}(S(\gamma^0)-S(\gamma^1))\}}.$$

Since $p(\boldsymbol{\beta}_\gamma|\boldsymbol{\lambda},\sigma^2)\propto p(\boldsymbol{\lambda}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\sigma^2)p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2)$,

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\lambda},\sigma^2)\propto\exp\{-\frac{1}{2}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}-\boldsymbol{\beta}_0)'V_{\boldsymbol{\gamma}_i}^{-1}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}-\boldsymbol{\beta}_0)\}.$$

So the posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\boldsymbol{\lambda},\sigma^2\sim MN(\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}'_{\boldsymbol{\gamma}}\boldsymbol{\lambda},\mathbf{V}_{\boldsymbol{\gamma}}).$$

Similarly the conditional distribution of σ^2 is again an inverse-gamma distribution.

It is clear from the expressions that due to the introduction of ϵ_i, $\boldsymbol{\beta}$ and σ and $\boldsymbol{\gamma}$ are conditionally independent of T and α, hence standard normal linear model (Lindley and Smith, 1978) results can be exploited to generate from the conditional

distributions or to marginalize them. We make use of these conditional distribution by simulating a Gibbs sampler that iterates through the following steps: (i) update θ; (ii) update $\boldsymbol{\lambda}$; (iii) update $\boldsymbol{\gamma}$, (iV) update $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

For the update to $\boldsymbol{\lambda}$, we propose to update each λ_i in turn conditional on the rest. That is, we update $\lambda_i|\boldsymbol{\lambda}_{-i}, \mathbf{y}, \theta, \sigma^2, \boldsymbol{\beta}$ $(i = 1, \ldots, n)$, where $\boldsymbol{\lambda}_{-i}$ indicates the $\boldsymbol{\lambda}$ vector with the ith element removed.

The conditional distribution of λ_i does not have an explicit form; we thus resort to the Metropolis-Hastings (MH) procedure with a proposal density $q(\lambda_i, \lambda_i^*)$ that generates moves from the current state λ_i to a new state λ_i^*. The proposed updates are then accepted with probabilities

$$\alpha = \min\left\{1, \frac{p(y_i|\lambda_i^*)p(\lambda_i^*|\boldsymbol{\lambda}_{-i}, \text{others})q(\lambda_i, \lambda_i^*)}{p(y_i|\lambda_i)p(\lambda_i|\boldsymbol{\lambda}_{-i}, \text{others})q(\lambda_i, \lambda_i^*)}\right\}; \tag{3.2}$$

otherwise the current state is retained. Similarly we need MH steps to draw θ.

3.3. Proportional Hazards Regression Model

The Cox proportional hazards model (Cox, 1972) assumes that the hazard function consists of two parts: baseline hazard function and nonnegative function of covariates. It is given by

$$h(t|\boldsymbol{x}) = h_0(t)\exp(W)$$

where $h_0(t)$ is the baseline hazard function and $W = \boldsymbol{x}'\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a vector of regression coefficients. The Weibull model in the previous section is a special case of Cox's proportional hazard model with $h_0(t) = \alpha t^{\alpha-1}$. Due to indetermination of baseline hazard function, the proportional hazards (PH) model has adequately adaptable for many applications (Kalbfleisch, 2002).

Kalbfleisch (1978) suggested the nonparametric Bayesian method for the PH

model. We apply Bayesian variable selection approach to this model. In addition, similar to Weibull regression model, we overcame the computation difficulties by including a random residual component as

$$W_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \text{ where } \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

where $\mathbf{X}$ is the design matrix with ith column $\boldsymbol{x}_i$.

Let T_i be a independent random variable with conditional survival function

$$P(T_i \geq t_i | W_i, \Lambda) = \exp\{-\Lambda(t_i)\exp(W_i)\} \qquad (i = 1, \cdots, n).$$

Kalbfleisch (1978) suggested a Gamma process ($\mathcal{GP}$) prior for the baseline cumulative hazard function. He supposed that $\Lambda \sim \mathcal{GP}(a\Lambda^*, a)$ where Λ^* is the mean process and a is a weight parameter about the mean (Ibrahim *et al.*, 2001). Kalbfleisch(1978) showed that if $a \approx 0$, the likelihood is approximately proportional to the partial likelihood and if $a \to \infty$, the limit of the likelihood is the same as the gamma process is replaced by Λ^*. Since $\Lambda(t) \sim \mathrm{Gam}(a\Lambda^*(t), a)$ for given t, the unconditional marginal survival function is obtained by direct integration:

$$\begin{aligned} P(T_i \geq t|W) &= \int \exp\left(-r\exp(W)\right) \frac{r^{a\Lambda^*(t)-1}}{\Gamma(a\Lambda^*(t))a^{-a\Lambda^*(t)}} \exp(-ar)dr \\ &= \left(\frac{a}{a+\exp(W)}\right)^{a\Lambda^*}. \end{aligned}$$

The joint survival function conditional on Λ is

$$P(T_1 \geq t_1, \cdots, T_n \geq t_n | \mathbf{W}, \Lambda) = \exp\{-\sum \Lambda(t_i)\exp(W_i)\}.$$

Using a property of Gamma process, Kalbfleisch(1978) showed that the likelihood

with some right censoring is

$$L(\mathbf{W}|\mathbf{D}) = \exp\{-\sum aB_i\Lambda^*(t_i)\}\prod_1^n\{a\lambda^*(t_i)B_i\}^{\nu_i}$$

where

$$\nu_i = \begin{cases} 0 & \text{if } t_i \text{ is right censored} \\ 1 & \text{if } t_i \text{ is a death time} \end{cases},$$

$A_i = \sum_{l\in R(t_i)}\exp(\mathbf{W}_l) \quad (j = 1, \cdots, n)$, $R(t_i)$ is the set of individuals at risk at time $t_i - 0$, $B_i = -\log\{1 - \exp(W_i)/(a + A_i)\}$ and $\mathbf{D} = (n, \boldsymbol{t}, \boldsymbol{\nu})$ denotes the observed data. The prior distributions are as follows:

$$\begin{aligned}
[\mathbf{W}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}] &\sim \mathrm{MN}(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2\mathrm{I}) \\
[\boldsymbol{\beta}_{\boldsymbol{\gamma}}] &\sim \mathrm{MN}(0, \mathrm{c}\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}) \\
[\gamma_i] &\sim \text{Bernoulli}(\pi_i) \\
[\sigma^2] &\sim \text{Inverse}\,\text{Gamma}\left(a_0, \frac{b_0}{2}\right).
\end{aligned}$$

All conditional distributions are obtained in the next section.

3.3.1. Conditional Distributions in the Cox's Proportional Model

The full conditional distribution of $\mathbf{W}$ is:

$$\begin{aligned}
p(\mathbf{W}|\mathbf{D}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2) \quad \propto \quad & \exp\{-\sum aB_i\Lambda^*(t_i)\}\prod_1^n\{a\lambda^*(t_i)B_i\}^{\nu_i} \\
& \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{W} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})'(\mathbf{W} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\right\}
\end{aligned}$$

In order to get the $p(\boldsymbol{\gamma}|\mathbf{W}, \sigma^2)$, we need to integrate out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and the approach is similar to the Weibull regression situation.

The marginal distribution of $\boldsymbol{\gamma}$ given $\mathbf{W}$ and σ^2 is

$$p(\boldsymbol{\gamma}|\mathbf{W}, \sigma^2) \quad \propto \quad p(\mathbf{W}|\boldsymbol{\gamma}, \sigma^2)p(\boldsymbol{\gamma})$$

$$\propto \quad \exp\left\{-\frac{1}{2\sigma^2}S(\boldsymbol{\gamma})\right\}\prod_{i=1}^{n}{\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i}$$

where $S(\boldsymbol{\gamma}) = \mathbf{W}'\mathbf{W} - \frac{c}{1+c}\mathbf{W}'\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{W}$. Now rather than drawing γ as a vector better to draw component wise from $p(\gamma_i|\lambda, \gamma_{j\neq i})$ which is

$$\begin{aligned}
p(\gamma_i|\mathbf{W}, \boldsymbol{\gamma}_{j\neq i}, \sigma^2) &\propto p(\mathbf{W}|\boldsymbol{\gamma}, \sigma^2)p(\gamma_i) \\
&\propto (1+c)^{-q\boldsymbol{\gamma}/2}\exp\left\{-\frac{1}{2\sigma^2}S(\boldsymbol{\gamma})\right\}{\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i} \\
p(\gamma_i = 1|\mathbf{W}, \boldsymbol{\gamma}_{j\neq i}, \sigma^2) &\propto \pi_i(1+c)^{-q{\boldsymbol{\gamma}_i}^1/2}\exp\left\{-\frac{1}{2\sigma^2}S({\boldsymbol{\gamma}_i}^1)\right\} \\
p(\gamma_i = 0|\mathbf{W}, \boldsymbol{\gamma}_{j\neq i}, \sigma^2) &\propto (1-\pi_i)(1+c)^{-q{\boldsymbol{\gamma}_i}^0/2}\exp\left\{-\frac{1}{2\sigma^2}S({\boldsymbol{\gamma}_i}^0))\right\}
\end{aligned}$$

where ${\boldsymbol{\gamma}_i}^1 = (\gamma_1, \cdots, \gamma_i = 1, \cdots, \gamma_p)$ and ${\boldsymbol{\gamma}_i}^0 = (\gamma_1, \cdots, \gamma_i = 0, \cdots, \gamma_p)$,

$$p(\gamma_i = 1|\mathbf{W}, \gamma_{j\neq i}, \sigma^2) = \frac{1}{1 + \frac{1-\pi_i}{\pi_i}(1+c)^{1/2}\exp\{-\frac{1}{2}(S(\gamma^0) - S(\gamma^1))\}}.$$

The conditional distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is again

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \mathbf{W}, \sigma^2 \sim MN(\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}'\mathbf{W}, \mathbf{V}_{\boldsymbol{\gamma}})$$

and the conditional distribution of σ^2 is Inverse-gamma.

3.3.2. Bayes Factor Computation

In this non-conjugate complicated model, derivation of the marginal likelihoods in explicit form is not possible. Chib and Jeliazkov (2001) overcame the problem of calculation of the marginal likelihood by estimating it through Metropolis–Hastings (M–H) output. We basically follow their two parameter blocks and multiple latent variable blocks algorithm to get the Bayes Factors for Weibull model and modify it for the Cox's proportional hazard model.

For the Weibull model, let $q(\boldsymbol{\lambda}, \boldsymbol{\lambda}'|\mathbf{D}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \theta)$ be the proposal density for the tran-

sition from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}'$. And let

$$\alpha(\boldsymbol{\lambda}, \boldsymbol{\lambda}'|\mathbf{D}, \boldsymbol{\beta\gamma}, \theta) = \min\left\{1, \frac{f(\mathbf{D}|\boldsymbol{\lambda}', \theta, \boldsymbol{\gamma})\pi(\boldsymbol{\lambda}', \theta)}{f(\mathbf{D}|\boldsymbol{\lambda}, \theta, \boldsymbol{\gamma})\pi(\boldsymbol{\lambda}, \theta)} \times \frac{q(\boldsymbol{\lambda}', \boldsymbol{\lambda}|D, \boldsymbol{\theta}, \boldsymbol{\beta\gamma})}{q(\boldsymbol{\lambda}, \boldsymbol{\lambda}'|D, \boldsymbol{\theta}, \boldsymbol{\beta\gamma})}\right\}$$

denote the acceptance probability for the move from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}'$. Using a basic marginal likelihood identity

$$m(\mathbf{D}) = \frac{f(\mathbf{D}|\boldsymbol{\lambda}^*, \theta^*)\pi(\boldsymbol{\lambda}^*, \theta^*)}{\pi(\boldsymbol{\lambda}^*, \theta^*|\mathbf{D})} \quad \text{for any } \boldsymbol{\lambda}^* \text{ and } \theta^*$$

and the following decomposition

$$\pi(\boldsymbol{\lambda}^*, \theta^*|\mathbf{D}) = \pi(\boldsymbol{\lambda}^*|\mathbf{D})\pi(\theta^*|\mathbf{D}, \boldsymbol{\lambda}^*),$$

the estimation of the marginal likelihood reduces to estimation of $\pi(\boldsymbol{\lambda}^*|\mathbf{D})$ and $\pi(\theta^*|\mathbf{D}, \boldsymbol{\lambda}^*)$. Let

$$p(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma}) = \alpha(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})$$

be the subkernel of the M-H chain for $\boldsymbol{\lambda}$ conditioned on $(\theta, \boldsymbol{\beta\gamma})$ By a local reversibility condition

$$p(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})\pi(\boldsymbol{\lambda}|\mathbf{D}, \theta, \boldsymbol{\beta\gamma}) = \pi(\boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})p(\boldsymbol{\lambda}^*, \boldsymbol{\lambda}|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})$$

and by integrating over $\boldsymbol{\psi} = (\boldsymbol{\lambda}, \theta, \boldsymbol{\beta\gamma})$ after multiplying both sides of the local reversibility condition by $\boldsymbol{\pi}(\theta, \boldsymbol{\beta\gamma}|\mathbf{D})$, we obtain finally

$$\boldsymbol{\pi}(\boldsymbol{\lambda}^*|\mathbf{D}) = \frac{E_1\left\{\alpha(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})\right\}}{E_2\left\{\alpha(\boldsymbol{\lambda}^*, \boldsymbol{\lambda}|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})\right\}}$$

where E_1 is expectation with respect to the distribution $\boldsymbol{\pi}(\boldsymbol{\lambda}, \theta, \boldsymbol{\beta\gamma}|\mathbf{D})$ and E_2 is expectation with respect to the distribution $\boldsymbol{\pi}(\theta, \boldsymbol{\beta\gamma}|\mathbf{D}, \boldsymbol{\lambda}^*) \times q(\boldsymbol{\lambda}^*, \boldsymbol{\lambda}|\mathbf{D}, \theta, \boldsymbol{\beta\gamma})$. The numerator can be estimated by averaging the full run Monte Carlo samples and gthe

denominator can be estimated by averaging the additional MCMC samples.

Since we cannot get the normalizing constant of $\pi(\boldsymbol{\theta}|\mathbf{D}, \boldsymbol{\lambda}^*)$, we need a similar step to estimate it.

$$\hat{\pi}(\theta^*|\mathbf{D}, \boldsymbol{\lambda}^*) = \frac{M^{-1}\sum_{g=1}^{M} \alpha(\theta^{(g)}, \theta^*|\mathbf{D}, \boldsymbol{\lambda}^*) q(\theta^{(g)}, \theta^*|\mathbf{D}, \boldsymbol{\lambda}^*)}{J^{-1}\sum_{j=1}^{J} \alpha(\theta^*, \theta^{(j)}|\mathbf{D}, \boldsymbol{\lambda}^*)}$$

where $\theta^{(g)}$ from $\pi(\theta|\mathbf{D}, \boldsymbol{\lambda}^*)$ and $\theta^{(j)}$ from $q(\theta^*, \theta|\mathbf{D}, \boldsymbol{\lambda}^*)$ After that, we can get the estimated logarithm of the marginal likelihood given by

$$\log \hat{m}(\mathbf{D}) = \log f(\mathbf{D}|\boldsymbol{\lambda}^*, \theta^*) + \log \pi(\boldsymbol{\lambda}^*, \theta^*) - \left\{\log \hat{\pi}(\boldsymbol{\lambda}^*|\mathbf{D}) + \log \hat{\pi}(\theta|\mathbf{D}, \boldsymbol{\lambda}^*)\right\}.$$

For the Cox's Proportional model, the estimation of the marginal likelihood is very similar to the method for the Weibull model. The basic marginal likelihood identity is

$$m(\mathbf{D}) = \frac{f(\mathbf{D}|\mathbf{W}^*)\pi(\mathbf{W}^*)}{\pi(\mathbf{W}^*|\mathbf{D})}.$$

An estimate of the marginal ordinate is similarly

$$\hat{\pi}(\mathbf{W}^*|\mathbf{D}) = \frac{M^{-1}\sum_{g=1}^{M} \alpha(\mathbf{W}^{(g)}, \mathbf{W}^*|\mathbf{D}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(g)}) q(\mathbf{W}^{(g)}, \mathbf{W}^*|\mathbf{D}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(g)})}{J^{-1}\sum_{j=1}^{J} \alpha(\mathbf{W}^*, \mathbf{W}^{(j)}|\mathbf{D}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)})}$$

where $\{\mathbf{W}^{(g)}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(g)}\}$ from the full run and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)}$ from $\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{D}, \mathbf{W}^*)$, $\mathbf{W}^{(j)}$ from $q(\mathbf{W}^*, \mathbf{W}|\mathbf{D}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(j)})$. Therefore, the logarithm marginal likelihood estimate is

$$\log \hat{m}(\mathbf{D}) = \log f(\mathbf{D}|\mathbf{W}^*) + \log \pi(\mathbf{W}^*) - \log \hat{\pi}(\mathbf{W}^*|\mathbf{D}).$$

We can use these marginal likelihood calculations to obtain the Bayes factor to compare two models. Kass and Raftery (1995) suggested that evidence against null model H_0 is very strong if twice the natural logarithm of the Bayes factor is greater than 10.

3.4. Examples

We apply the proposed methods to two data sets: Diffuse Large B-Cell Lymphoma Data (Alizadeh *et al.*, 2000) and Breast Carcinoma Data (Sørlie *et al.*, 2001). In all the examples we used MCMC method with a burn-in of 10,000 samples, after which every 10th sample was retained in the next 100,000 samples.

3.4.1. DLBCL Data

We applied these methods for finding a set of responsible genes which explain the survival function to a Diffuse Large B-cell lymphoma (DLBCL) data set (Alizadeh *et al.*, 2000). Diffuse large B-cell lymphoma (DLBCL) is one of subtypes of non-Hodgkin's lymphoma. But still patients with this disease had diverse responses to current therapy. So Alizadeh *et al.* (2000) proposed that there should be some different forms of DLBCL and discovered two distinct forms of DLBCL, activated B-like DLBCL and GC–B like, using DNA microarray experiment and hierarchical clustering. They showed that these two subgroups of DLBCL were differentiated from each other by distinct gene expressions of hundreds of different genes and had different survival time patterns.

There are 40 patients and expression level measurements for 4513 genes for each patient. We consider the fixed binary covariate X_0 as $X_{i0} = 1$ if the ith sample is Activated B-like and $X_{i0} = 0$ if other case for $i = 1, \cdots, 40$. Also we have the expression level measurement for a set of genes, so X_{ij} is the normalized log scale measurement of the expression level of the jth gene for the ith sample, where $i = 1, \cdots, 40$ and $j = 1, \cdots, 4513$.

3.4.1.1. Weibull Model

First we have used the parametric Weibull model. Using a two-sample t-test, 1000 genes are preselected and for the initial value of γ, the top five genes with largest absolute t–value are used. The best subset found by the MCMC chains had only 4 genes included. We fixed this model and and reran the MCMC for additional draws of $\{\boldsymbol{\lambda}^{(g)}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(g)}, \theta^{(g)}\}_{g=1}^{G}$. The survival function of Weibull model is $S(t|\lambda, \alpha) = \exp\left(-\exp(\lambda)t^{e^{\theta}}\right)$ and using the MCMC samples we obtain its Monte Carlo estimates:

$$\hat{S}_k(t|\lambda, \alpha) = \exp\left(-\exp(\bar{\lambda}_k)t^{e^{\bar{\theta}}}\right),\ k = 1 \text{ or } 2 \tag{3.3}$$

where k indicates class, $\bar{\lambda}_k$ is the MCMC samle mean of λ in the class k and $\bar{\theta}$ is the MCMC sample mean of θ. The 5th and 95th estimates of survival function use the 5th MCMC sample percentile of λ and 95th estimates of survival function use the 5th and 95th MCMC sample percentile of λ respectively instead of the MCMC sample mean in equation 3.3. Also we compare this model with the no gene model (model only with covariate X_0) and the $2\log(\text{Bayes Factor})$ comes out to be 33.98 which shows strong support for the selected model. Figure 4 shows two superimposed survival curves based on Kaplan–Meier method (dash–dotted line) and Weibull model (solid line) with 5th and 95th line (dotted line) for two groups, GC B–like (red) and Activated B–like (blue). Comparing to the Kaplan–Meier curve, it is clear that the fits are not satisfactory so we consider the semiparametric Cox's proportional hazard model to improve it.

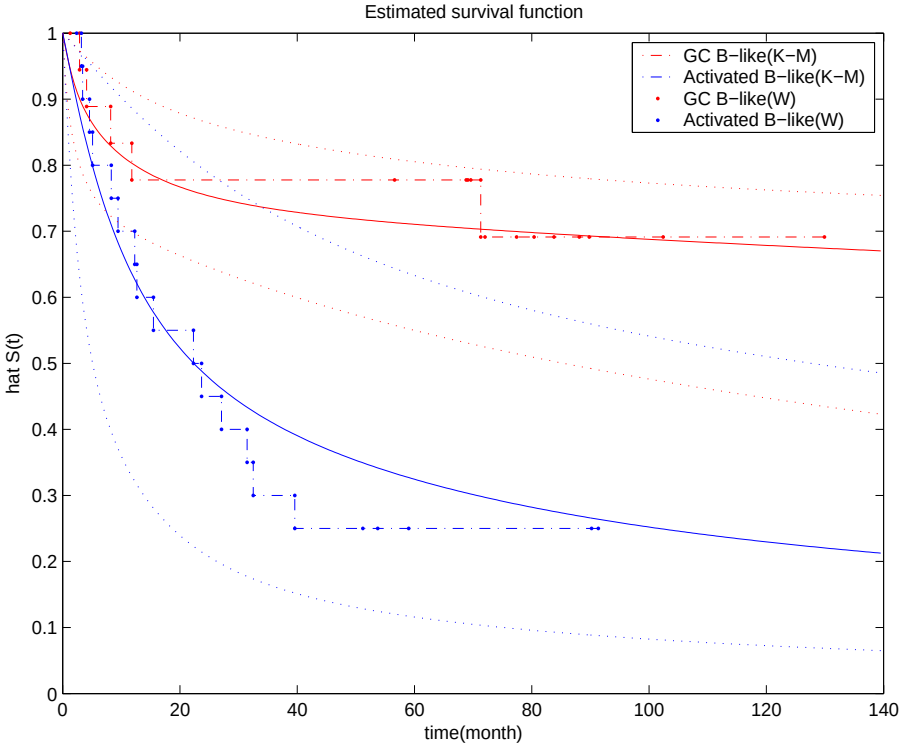

Fig. 4. Survival Function for DLBCL Data Using Weibull Model

3.4.1.2. Cox's Proportional Hazards Model

To develop the semiparametric model, we choose the baseline function Λ^* as Weibull distribution for the Gamma process on cumulative hazard function, that is, $\Lambda^*(t) = \eta_0 t^{\kappa_0}$. We choose the hyperparameter as $a = 10000$. The estimates of hyperparameters, η_0 and κ_0, are obtained using estimates of intercept and scale in Survreg function (survReg(formula=Surv(y, censor) 1, dist="weibull") in S+. For our computational convenience, 1000 genes are preselected by a two-sample t-test.

A 5 gene model comes out to be the best subset from the MCMC chain. We compared this model with the best Weibull model using the Bayes factor and 2 log(Bayes Factor) value is 14.81, which shows strong support for the semiparametric model. The survival function in the Cox's proportional hazard model is

$$S(t|W) = P(T \geq t|W) = \left(\frac{a}{a + \exp(W)}\right)^{a\Lambda^*}, \tag{3.4}$$

and we exploited the posterior samples for this model to get the Monte Carlo estimate of the function:

$$\hat{S}_k(t|W) = P(T \geq t|W) = \frac{1}{n_k} \sum_{l \in G_k} \left(\frac{a}{a + \exp(\bar{W}_l)}\right)^{a\Lambda^*} \text{ for } k = 1 \text{ or } 2.$$

where $\bar{W}_l$ is the MCMC sample mean of W_l, G_k is the group of samples in class k and n_k is the size of G_k. The 5th and 95th survial function estimates use 5th percentile and 95th percentil MCMC samples respectively instead of MCMC sample mean in equation 3.4. The posterior estimates of survival curves (solid line) with 5th and 95th survival estimates(dotted line) are superimposed on the Kaplan-Meier estimates (dash-dotted line) of survival functions (Figure 5). These plots show that this model is a good fit to both of the subgroup of patients.

Rather than a single, parsimonious model, the biologists may be interested in

bigger families of genes to study relationships and functions. We presented some selected genes based on the marginal frequencies in Table IX. Some of the identified genes are already known to be biologically significant. Since MAPK10 (mitogen-activated protein kinase 10) is connected to TNF(tumor necrosis factor)-a signaling pathway(Decraene *et al.* 2002), its expression is directly related to the existence of tumor. Rimokh *et al.* (1993) showed that FVT1 (follicular variant-translocation gene) is highly expressed in some T-cell malignancies. WASIP(Wiskott-Aldrich syndrome protein-interacting protein) is known to play a role in cortical actin assembly related to lymphocyte function according to Ramesh *et al.* (1997).

A heatmap based on top two genes in Figure 6 shows that these two gene expression patterns are related to survival time and are distinct between two groups. That is, GC B-like and Activated B-like have different gene expressions and survival times.

Table IX. Responsible Genes Found for Estimating the Survival Function DLBCL Data

Freq	Clone ID	Gene Symbol	Gene Name
1014	1355868		
910	290230	ICSBP1	interferon consensus sequence binding protein 1
405	814260	FVT1	follicular lymphoma variant translocation 1
289	1353111	MDS019	phorbolin-like protein MDS019
180	683069		EST
156	1340233	FGD3	"FGD1 family, member 3"
156	23173	MAPK10	mitogen-activated protein kinase 10
154	814601		
132	1335070		ESTs
132	1303587		
124	1337701	WASPIP	Wiskott-Aldrich syndrome protein interacting protein
121	824198	(FUBP1)	Homo sapiens far upstream element (FUSE) binding protein 1

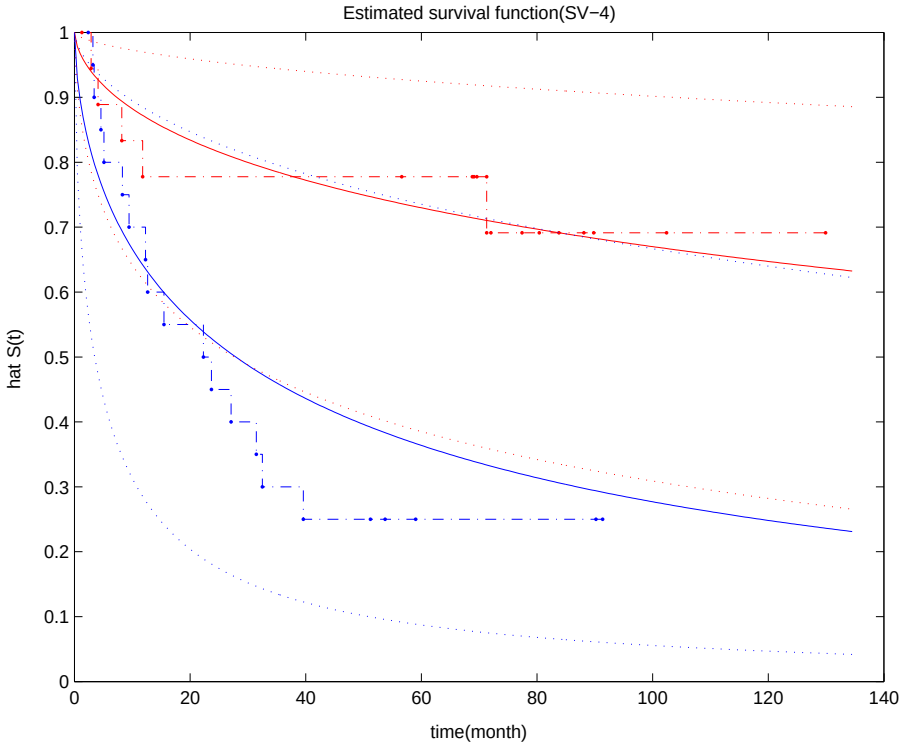

Fig. 5. Survival Function for DLBCL Data Using Semiparametric Hazards Model

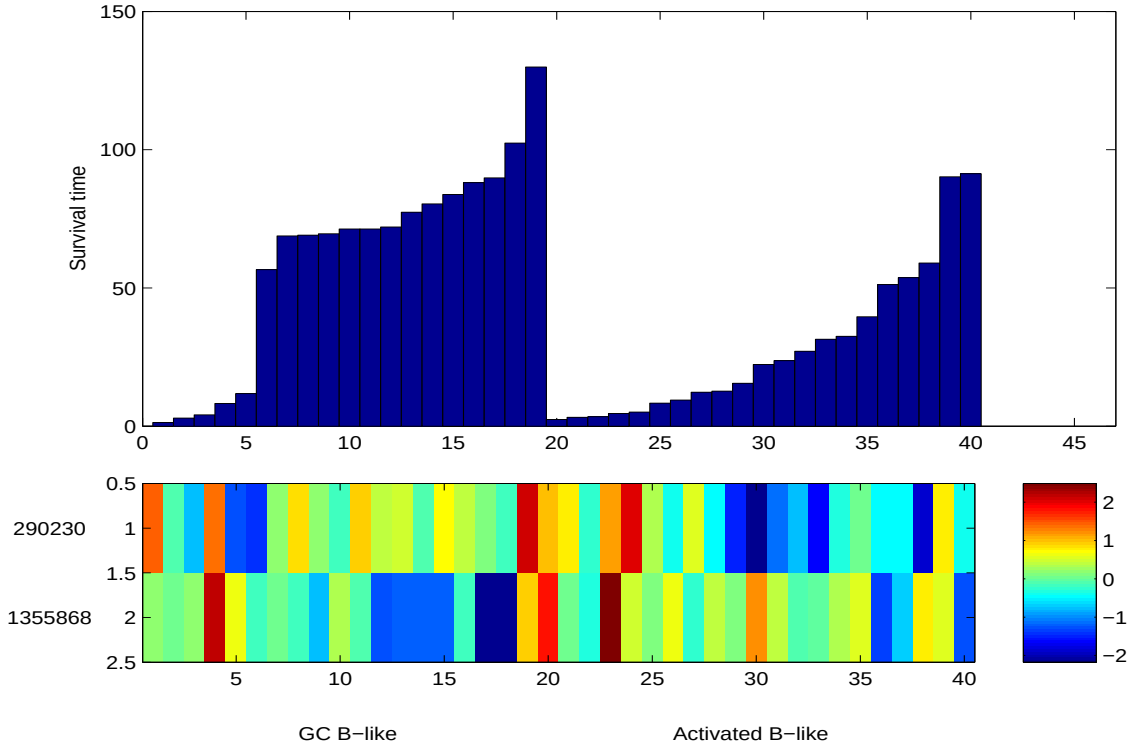

Fig. 6. Heat Map with Survival Time for DLBCL Data

3.4.2. Breast Carcinoma Data Set

Breast carcinoma data was described in Sørlie *et al.* (2001). They found some novel subclasses of breast carcinoma based on gene expression patterns and proved robustness of subclasses using separate gene sets. A part of patients with the same therapy had different overall and relapse-free survival patterns. We use overall survival times of 76 samples each with 3097 gene expressions. There is also additional subgroup information: Basal-like, ERBB2+, Normal Breast-like, Luminal Subtype A, B, or C. We only consider the binary covariate X_0 as: $X_{i0} = 1$ if ith sample is Luminal subtype A, B, or C and 0 otherwise. We can extend it to a muti-category case by

adding other covariates in the model.

3.4.2.1. Weibull Model

The best subset found by the MCMC chain had only 6 genes included. The 2 log (Bayes factor) of this model compared to the no-gene model was calculated to be 33.46. The posterior survival curve has been plotted in Figure 7, which again shows there is still room to improve the fitness of the model. Figure 7 shows two superimposed survival curves based on Kaplan-Meier method (dash-dotted line) and Weibull model (solid line) with 5th and 95th line (dotted line) for two groups, Luminal subtyes A,B, or C (red) and others (blue). Comparing to the Kaplan-Meier curve, it is clear that the fits are not satisfactory so we consider the semiparametric Cox's proportional hazard model to improve it.

3.4.2.2. Cox's Proportional Model

We perform our analysis with a semiparametric model using the previous data. A 5 gene model becomes the best subset. The 2 *log* Bayes factor value compared to the best Weibull model is 19.38. The posterior survival curves based on this model is given in Figure 8. Figure 8 shows two superimposed survival curves based on Kaplan-Meier method (dash-dotted line) and Cox's proportional hazard model (solid line) with 5th and 95th line (dotted line) for two groups, Luminal subtyes A,B, or C (red) and others (blue).

We presented some of the selected genes based on the marginal frequencies in Table X. Some of the genes are already shown to significant as Theillet *et al.* (1993) showed that PLAT(plasminogen activator) could be amplified in breast cancer. CYP1B1(cytochrome p450-1B1) plays a possibly critical role in the cause of breast cancer (Zheng *et al.* 2000).

A heatmap of these genes is provided in Figure 9 and there are some patterns: selected genes are overexpressed in group1 and survival times are related to the level of gene expressions. Three genes are selected through our proposed method. Two survival groups show different patterns and this finding corresponds to the pattern of actual gene expressions.

3.5. Discussion

We have proposed Bayesian models for variable selection in the survival regression models with specific application to analyze microarray data. We obtain nice estimate of the survival curves with an extremely small number of genes. On the other hand, bigger families of genes can be useful to biologists to study the relationship and functions. Information on the size of models for prediction can be easily included in our Bayesian search of good models. The method has flexbility of allowing the location of larger sets of genes, via the inspection of the best visited models or the marginal probabilities of single genes, as we have illustrated.

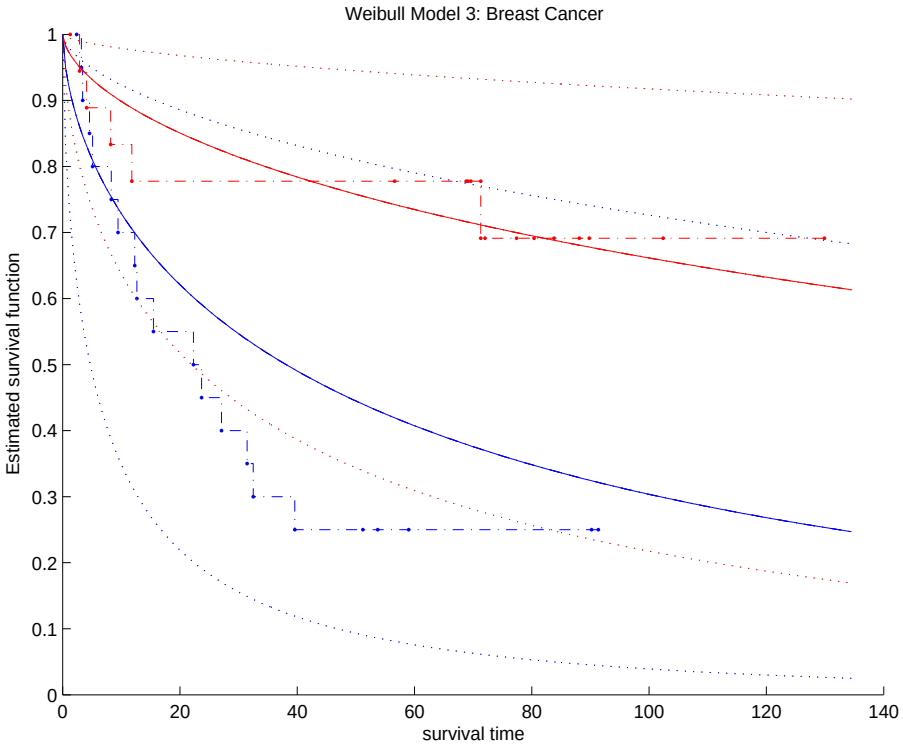

Fig. 7. Survival Function for Breast Carcinoma Data Using Weibull Model

Table X. Responsible Genes Found for Estimating the Survival Function for Breast Carcinoma Data

Freq.	Clone ID	Gene Symbol	Gene Description
3826	340826		Homo sapiens cDNA FLJ12749 fis, clone NT2RP2001149
1225	772890		Homo sapiens mRNA; cDNA DKFZp434N2412
1203	272018		ESTs, Weakly similar to AF126743 ...
1028	772304	SLC25A5	solute carrier family 25
593	813841	PLAT	plasminogen activator, tissue
573	745402	PRO2975	hypothetical protein PRO2975
396	40299	GDF10	growth differentiation factor 10
395	79045	CTL1	transporter-like protein
365	950578	NDUFA5	NADH dehydrogenase 1 alpha subcomplex, 5
263	782760	CYP1B1	cytochrome P450, subfamily I, polypeptide 1
229	768299	BRF1	butyrate response factor 1 (EGF-response factor 1)
198	951125	PECI	peroxisomal D3,D2-enoyl-CoA isomerase
185	131653	MRPS12	mitochondrial ribosomal protein S12
151	771323	PLOD	procollagen-lysine, 2-oxoglutarate 5-dioxygenase
143	755599	IFITM1	interferon induced transmembrane protein 1 (9-27)
132	214448	RNF10	ring finger protein 10
113	83610	FLJ22378	hypothetical protein FLJ22378
111	134476	SYBL1	synaptobrevin-like 1

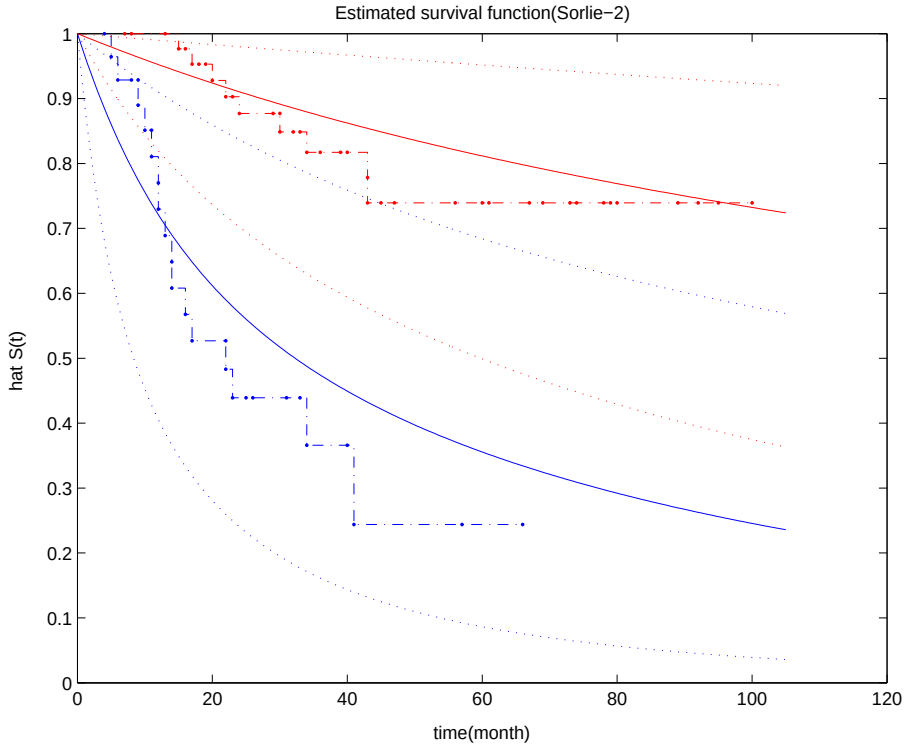

Fig. 8. Survival Function for Breast Carcinoma Data Using Semiparametric Model

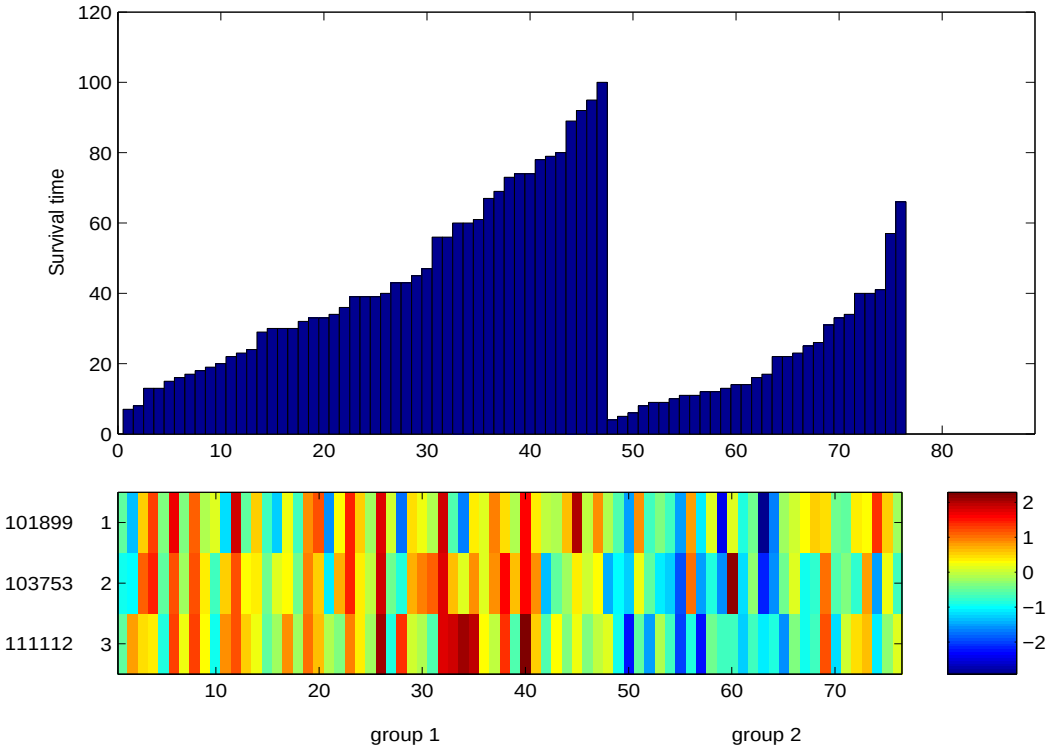

Fig. 9. Heat Map with Survival Time for Breast Carcinoma Data

CHAPTER IV

CURVE CLUSTERING IN THE MICROARRAY DATA

4.1. Introduction

After a novel technology, microarray, was introduced, many statistical issues have been raising including clustering. When gene expressions are observed by time or temperature, we want to cluster these gene expressions. In the setting of time–course gene expression, clustering methods can be divided into two categories: non–model–based methods and model–based methods. Non–model–based methods are such as hierarchical clustering (Eisen *et al.*,1998), clustering using correlation(Chu *et al.*,1999), self–organizing maps (Tamayo *et al.*,1999). Mixture-effects model with B–splines (Luan *et al.*,2003) and hidden Markov model (Schliep *et al.*,2003) are considered methods in model–based methods. We are motivated by Wakefield *et al.*(2003) who modeled the trajectory as a function of time and gene specific parameters and clustered these curves based on gene specific parameter using a reversible jump MCMC. Wakefield *et al.*(2003) used a first–order random walk model for gene–based parameters in a sporulation data (Chu *et al.*, 1999) and a mixture of periodic function model for the cell–cycle data (Spellman *et al.* 1998).

In this chapter, we propose a mixture of Dirichlet processes model using discrete wavelet transform for curve clustering as a fully Bayesian approach. In order to characterize these time–course gene expressions, we consider them as trajectory functions of time and gene specific parameters and obtain their wavelet coefficients by discrete wavelet transform. We then build cluster curves based using a mixture of Dirichlet processes prior. Each iteration of MCMC algorithm generates the cluster structure of these coefficients as a by–product (Escobar *et al.,* 1998). In addition, when mi-

croarray have missing data points, we estimate their missing data points in MCMC sampling using their conditional distribution. Subsequently, the proposed models are applied to two yeast cell cycle microarray data sets: Cho *et al.* (1998) and Spellman *et al.* (1998).

4.2. Bayesian Hierarchical Model

4.2.1. Mixture of Dirichlet Processes

Ferguson(1973) defined a Dirichlet process ($\mathcal{DP}$) for a Bayesian nonparametrci approach as follows: Let μ be a finite non-null measure on $(\mathcal{X}, \mathcal{B})$ where $\mathcal{X}$ is a space and $\mathcal{B}$ is a σ-field of subsets. If for any $k \in \{1, 2, \cdots, \}$ and any measurable partition $(B_1, \cdots, B_k)$ of $\mathcal{X}$,

$$(P(B_1), \cdots, P(B_k)) \sim \text{Dirichlet}(\mu(B_1), \cdots, \mu(B_k))$$

then a stochastic process P is defined as a Dirichlet process $\mathcal{DP}(\mu)$ on $(\mathcal{X}, \mathcal{B})$ with parameter μ. Especially, when $\mu = \alpha G_0$ with G_0 is distribution, $E(P(B)) = G_0(B)$ and $Var(P(B)) = \dfrac{G_0(B)(1 - G_0(B))}{\alpha + 1}$ for any $B \in \mathcal{B}$. In this case,G_0 is called the base measure, α is called the precision parameter, and $\mathcal{DP}$ is denoted by $\mathcal{DP}(\alpha, G_0)$ to acknowledge the dependence μ through α and G_0. Escobar and West(1998) explored Dirichlet process $\mathcal{DP}(\alpha, G_0)$ in order to model the "uncertainty" of the prior distribution, while referring to G_0 as the "location" distribution of Dirichlet process prior. We are interested in the following property of Dirichlet process: given a set of $\boldsymbol{\beta} = \{\beta_1, \cdots, \beta_I\}$ from a random distribution G following a Dirichlet process $\mathcal{DP}(\alpha, G_0)$, the conditional distribution

$$\beta_i | \boldsymbol{\beta}_{-i} \sim \frac{\alpha}{\alpha + I - 1} G_0 + \frac{1}{\alpha + I - 1} \sum_{j \neq i} \delta(\beta_i | \beta_j), \quad \boldsymbol{\beta}_{-i} = \{\beta_j \in \boldsymbol{\beta} : j \neq i\}$$

follows a mixture of Dirichlet process. Formal definition of a mixture of Dirichlet Process can be found in Antoniak(1974).

An important property in the MDP model is that with positive probability some of the $\boldsymbol{\beta}_i$ have the same value because of the discretness of random measure of MDP (MacEachern *et al.*, 1998) and clustered property of data. Escobar and West(1998) point out the Polya urn representation for the joint posterior distribution of $[\boldsymbol{\beta}_i|\cdot]$

$$[d\boldsymbol{\beta}_i|\mathbf{Y}, \boldsymbol{\beta}_{-i}, \cdot] \propto \prod_{i=1}^{I} f(\mathbf{Y}_i|\boldsymbol{\beta}_i, \cdot)\frac{\alpha G_0(d\boldsymbol{\beta}_i|\sigma^2, \mathbf{V}) + \sum_{k\neq i}\delta(\boldsymbol{\beta}_i|\boldsymbol{\beta}_k)}{\alpha + i - 1}$$

4.2.2. Wavelet Regression

For the ith gene, Y_{it} is the normalized log-ratio of mRNA gene expression level relative to the gene expression of the reference cell at time t, where $i \in \{1, \cdots, I\}$ and $t \in \{1, \cdots, T\}$; I is the number of genes and T is total number of equally spaced time points. We assume that $\mathbf{Y}_i = (Y_i, \cdots, Y_{iT})$ is the the vector of observations of the ith trajectory function with additive white noise

$$Y_{it} = f(\boldsymbol{\theta}_i, t) + \epsilon_{it}, \epsilon_{it} \sim N(0, \sigma^2)$$

where $f(\boldsymbol{\theta}_i, t)$ is a trajectory function of a gene-specific set of parameter,$\boldsymbol{\theta}_i$ and time t (Wakefield *et al.*, 2003). The trajectory function can be represented in terms of shifted and dilated scale functions $\{\psi(t)\}$ and wavelet functions $\{\phi(t)\}$ as follows:

$$f(t) \approx \sum_{k=0}^{2^{j_0}-1} s_k\phi_{j_0k}(t) + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} w_{jk}\psi_{jk}(t)$$

where $J = \log_2 T$, $j \geq \jmath_0 \geq 0$, $u_k =< f, \phi_{j_0k} >$ and $w_{jk} =< f, \psi_{jk} >$ (Daubechies, 1992) or equivalently we may write the model as

$$f = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is the wavelet representation of the true function f and $\mathbf{X}'$ is the $T \times T$ orthogonal wavelet transformation. In this chapter, we only assume orthogonal wavelet bases and avoid more general representations for questions of stability and bias introduced in estimation.

4.2.3. Generic Wavelet Based Dirichlet Process Model

The proposed method looks for relevant clusters in the observed curves by the posterior sampling of the wavelet coefficients in Dirichlet process mixtures $\mathcal{DP}(\alpha, G_0)$.

The prior of covariance Σ is modified as in an example of normal structure in Escobar and West(1998) and assume the following hierarchical structure :

$$\begin{aligned}
[\mathbf{Y}_i|\boldsymbol{\beta}_i, \sigma^2] &\sim N(\mathbf{X}\boldsymbol{\beta}_i, \sigma^2\mathcal{I}), \\
[\boldsymbol{\beta}_i|\boldsymbol{\mu}, \Sigma] &\sim \mathcal{DP}\left(\alpha, MN(\boldsymbol{\mu}, \sigma^2\Sigma)\right), \\
[\sigma^2] &\sim IG\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right), \\
[\alpha] &\sim G(a, b).
\end{aligned}$$

$\Sigma = diag\left(\{v_{jk}\}, 0 \leq k \leq 2^{j-1}, j_0 \leq j \leq J\right)$and is intended for shrinkage with

$$[v_{jk}] \sim IG\left(\frac{s_{jk}}{2}, \frac{r_{jk}}{2}\right), \quad 0 \leq k \leq 2^{j-1}, j_0 \leq j \leq J$$

where $r_{j\cdot}$ and $s_{j\cdot}$ are specified levelwise to maintain a mean of roughly $n2^{-cj}$ for some constant c. Here, the constant c models the decay in the average size of wavelet coefficients and, thus, the mean of the inverse gamma prior, $\mathrm{E}(v_{jk}) = s_{jk}/(r_{jk} - 2)$, $r_{jk} > 2$ is specified to match this decay. For all k, fixing $r_{j\cdot} = c+2$, we get $s_{j\cdot} = cn2^{-cj}$.

The posterior distributions are as follows:

$$[\boldsymbol{\beta}_i|\mathbf{Y}, \{\boldsymbol{\beta}_k, k \neq i\}, \sigma^2, \Sigma] \propto \exp\left\{-\frac{\sigma^2}{2}\sum_{i=1}^{I}(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)'(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i))\right\}$$

$$\begin{aligned} & \times \left(\frac{\alpha}{\alpha + I - 1} MN(\boldsymbol{\mu}, \sigma^2 \Sigma) + \frac{1}{\alpha + I - 1} \sum_{j \neq i} \delta(\beta_i | \beta_j)_j \right) \\ \propto \quad & q_0 MN(\mu_i, \sigma^2 \mathbf{V}) + \sum_{j \neq i} q_j \delta(\beta_i | \beta_j) \end{aligned}$$

where $V = (\Sigma^{-1} + \mathcal{I})^{-1}, \mu_i = V(\Sigma^{-1}\mu + \mathbf{X'Y}_i)$ and the weights q_j are defined as

$$\begin{aligned} q_0 \quad & \propto \quad \alpha \phi(\mathbf{Y}_i | \mathbf{X}\boldsymbol{\mu}, \sigma^2(\mathcal{I} + \mathbf{X'}\Sigma\mathbf{X})) \\ q_k \quad & \propto \quad \phi(\mathbf{Y}_i | \mathbf{X}\boldsymbol{\beta}_k, \sigma^2 \mathcal{I}) \end{aligned}$$

subject to $\sum_{j \neq i} q_j = 1$, where $\phi(y|\theta, \Upsilon)$ is the multinormal density function of mean θ and covariance Υ. Since the conditional probability of sampling a new β is proportional to q_0, if it is small relative to the sum of other q_j's, the number of distinct β_i's is also small and samples of β's change much. Let superscript * denote distinct values. Escobar and West (1998) used a "remixing algorithm" in order to avoid this problem by resampling β_j^* at each iteration, and to, additionally, improve the convergence.

$$[\boldsymbol{\beta}_j^* | \mathbf{Y}, \sigma^2, \Sigma] \propto MN(\mu_j^*, \sigma^2 \mathbf{V}_j^*) \, \text{for each j} = 1, \cdots, \text{I}^*$$

where $V_j^* = (\Sigma^{-1} + |J(j)|\mathcal{I})^{-1}$, $\mu_j^* = V_j^*(\Sigma^{-1}\mu + \sum_{j \in J(j)} \mathbf{X'Y}_j)$ and $J(j)$ is the index set of j th cluster

Since

$$\begin{aligned} \left[\sigma^2, \boldsymbol{\beta} | \mathbf{Y}\right] \quad \propto \quad & \left(\frac{1}{2\sigma^2}\right)^{I \cdot T/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{I} (\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i)'(\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i) \right\} \\ & \times \left(\frac{1}{2\sigma^2}\right)^{I \cdot T/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{I} (\boldsymbol{\beta}_i - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\mu}) \right\} \\ & \times \text{IG}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right), \end{aligned}$$

the full conditional distribution of σ^2 after integrating out β is

$$\left[\sigma^2|\cdot\right] \sim \text{IG}\left(\frac{\nu_1 + N}{2}, \frac{S}{2}\right)$$

where $\mathbf{S} = \sum_{i=1}^{I} \left(\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \mathbf{Y}_i'\mathbf{Y}_i - \boldsymbol{\mu}_i'\mathbf{V}^{-1}\boldsymbol{\mu}_i\right) + \nu_2$ and $N = I \cdot T$.

In addition, with $(\beta_i|\Sigma, \sigma^2) \sim \text{N}(\mu, \sigma^2\Sigma)$, the posterior distribution of scaling parameters v_{jk} are drawn as

$$(v_{jk}|\beta_i, \sigma^2) \sim \text{IG}\left(\frac{s_{jk}^*}{2}, \frac{r_{jk}^*}{2}\right)$$

where $s_{jk}^* = I + s_{jk}$ and $r_{jk}^* = (\sigma^2)^{-1}\sum_{i=1}^{I}(\beta_{ik} - u_k)^2 + r_{jk}$.

The precision parameter α in the Dirichlet process plays an important role in determining the number of clusters. Assuming a continuous prior density for $p(\alpha)$, Escobar and West(1995) provided a distribution of number of components through Antoniak(1974)'s results

$$p(I^*|\alpha, I) = c_I(I^*)I!\alpha^{I^*}\Gamma(\alpha)/\Gamma(\alpha + I), \quad I^* = 1, \cdots, I.$$

where $c_I = p(I^*|\alpha = 1, I)$ and $\Gamma(\cdot, \cdot)$ is the Gamma function. According to the relationship between the Gamma function and the Beta function,

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha + I)} = \frac{(\alpha + I)\beta(\alpha + 1, I)}{\alpha\Gamma(I)},$$

where $\beta(\cdot, \cdot)$ is the Beta function, the $p(\alpha|I^*)$ can be written as follows:

$$\begin{aligned} p(\alpha|I^*) &\propto p(I^*|\alpha)p(\alpha) \\ &\propto p(\alpha)\alpha^{I^*-1}(\alpha + I)\int_0^1 \eta^\alpha(1 - \eta)^{I-1}d\eta \end{aligned}$$

and it can be considered as the marginal distribution (of α) from a joint distribution

for α and a latent variable η such that

$$p(\alpha, \eta | I^*) \propto p(\alpha)\alpha^{I^*-1}(\alpha + I)\eta^{\alpha}(1-\eta)^{I-1}.$$

Therefore choosing $p(\alpha)$ to be $\mathrm{G}(a, b)$, leads to

$$p(\alpha | I^*, \eta) \sim \pi_\eta \mathrm{G}(a + I^*, b - \log(\eta)) + (1 - \pi_\eta)\mathrm{G}(a + I^* - 1, b - \log(\eta)),$$

where

$$\frac{\pi_\eta}{1-\pi_\eta} = \frac{a + I^* - 1}{I(b - \log(\eta))}.$$

Next, η is updated as

$$p(\eta | \alpha, I^*) \propto \eta^{\alpha}(1-\eta)^{I-1} = Beta(\alpha + 1, I).$$

4.3. Missing Data Case

The gene expressions $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ are recorded expression levels at successive times points $(1, 2, \ldots, T)$. In case, the expressions at some time points are missing, we proceed as follows.

Recall that for the ith curve

$$(Y_i | \mu, \Sigma, \sigma^2) \sim \mathrm{N}(X\boldsymbol{\mu}, \sigma^2(I + X\Sigma X')), \ \forall i = 1 \ldots I$$

Stratify Y_i' as $(Y_{i(1)}', Y_{i(2)}')$ where $Y_{i(1)}$ is a vector of $r_i \times 1$ observations and $Y_{i(2)}$ is a vector of $(T - r_i) \times 1$ missing data. Write, $\Lambda = (\mathcal{I} + X\Sigma X')$, then corresponding to the split in the observed vector, create partitions in

$$X\mu = \begin{pmatrix} (X\mu_i)_1 \\ (X\mu_i)_2 \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \Lambda_{i11} & \Lambda_{i21} \\ \Lambda_{i21} & \Lambda_{i22} \end{pmatrix}$$

then the missing data is drawn conditional on the observed time points of the i^{th}

curve as

$$(Y_{i(2)}|Y_{i(1)},\cdot) \sim \mathrm{N}((X\mu_i)_2 + \Lambda_{i21}\Lambda_{i11}^{-1}(Y_{i(1)} - (X\mu_i)_1), \sigma^2(\Lambda_{i22} - \Lambda_{i21}\Lambda_{i11}^{-1}\Lambda_{i12}))$$

This extra sampling step is iterated with the Gibbs sampler in the previous section.

4.4. Application to cDNA Microarray Data

We apply our proposed hierarchical model to two yeast cell cycle data and check the model adequacy using the Bayesian Information Criterion, BIC, introduced by Schwarz (1978),

$$\mathrm{BIC} = -2 \cdot log\hat{L} + \log(n) \cdot p$$

where $\hat{L}$ is the maximized likelihood and p is the number of parameter in the model. If the ratio of BIC of two models is considered, it is known to provide an approximation of $2\log$(Bayes Factor) for sufficiently large sample n (Schwarz, 1978).

4.4.1. The Yeast Cell Cycle Data I

Cho *et al.* (1998) obtained time courses of more than 6000 genes over nearly two full cell cycles (17 time points) and found that 416 among them have periodic-fluctuations in the gene expression due to their cell cycle-dependency. Genes were categorized into five-phases, early G1, late G1, S, G2 and M by their different peak points. Cho *et al.* (1998) identified that 33 of 416 genes have peak points in two different phases. So we used 384 genes for application to our proposed model and compare our clustering results with their identified phases. The data was log transformed and standardized across the cell cycle.

We used the last 16 time points for the computational convenience in the Discrete Wavelet Transform. We compared our clustering result with Cho *et al.* (1998) using

the adjusted Rand index(Hubert and Arabie, 1985), which evaluates the measure of agreement of two different partitions of one data set and overcomes the problem of non-constant expected value of Rand index (Rand, 1971), the fraction of agreement. Especially Milligan and Cooper (1986) recommended the adjusted Rand index as an external measure after extensive comparisons. If the cluster is random, the expected value of adjusted Rand index is 0. Its maximum value is 1 which indicates the perfect match between the clustering result and the external standard. The more details on the adjusted Rand index is referred to Yeung and Ruzzo (2000).

Table XI shows the comparison of two partitions. The adjusted Rand index based on Table XI is 0.4563 and we can see how sharply they are clustered. BIC of the model (=2294.8) is much lower than the BIC of Cho *et al.*'s clustering (=3659.7). Cho *et al.* (1998) classified these genes based on their peak time but our proposed model is based on the trajectory pattern. So it may be the reason of relative low adjusted Rand index and the lower BIC of our model supports that our proposed model clusters these curves sharply. In addition, the distribution of I^* do not change much in the same analysis with various priors of α; $Gamma(0.0001, 0.0001)$, $Gamma(1, 1)$, $Gamma(2, 1)$, and so on.

Figure 10 shows five curve clusters of 384 genes by Cho *et al.* (1998) and Figure 11 shows those generated by our proposed model.Compared with Figure 10, our model shows clearer classification schemes ($\mathcal{C}1 - \mathcal{D}3, \mathcal{C}2 - \mathcal{D}1, \mathcal{C}3 - \mathcal{D}5, \mathcal{C}4 - \mathcal{D}2, \mathcal{C}5 - \mathcal{D}4$).

Table XI. Two Partitions of Yeast Cell Cycle Data ($\mathcal{C}$: Clusters by Cho *et al.* (1998) and $\mathcal{D}$: Clusters by Our Proposed Model)

Class	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_3$	Sums
$\mathcal{D}_1$	40	14	0	0	1	55
$\mathcal{D}_2$	7	117	0	0	38	162
$\mathcal{D}_5$	0	1	41	3	35	80
$\mathcal{D}_4$	3	0	7	37	1	48
$\mathcal{D}_3$	17	3	4	15	0	39
Sums	67	135	52	55	75	

4.4.2. The Yeast Cell Cycle Data II

Spellman *et al.* (1998) carried out DNA microarray experiments in order to make a new comprehensive list of yeast genes whose gene transcription levels have periodical patterns within the cell cycle. They generated three microarray data sets from the yeast cell cultures synchronized by different methods: α factor arrest, elutriation and arrest of a *cdc*15 temperature-sensitive mutant. Samples in the first data were taken every 7 minitures from 0 minture to 140 minute, but α factor was removed after 120 minutes. So the number of time points is 18. We pre–selected 400 genes with large variance over time. We transform the data using the Box–Cox family of power transformation, with $\lambda = 0.1$, for the normality.

$$Y_\lambda = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ log(Y), & \text{if } \lambda = 0 \end{cases}$$

We consider the missing data points and at every 5th iteration, the missing points are estimated to improve the speed of MCMC. Figure 12 shows that the estimated data does not show difference from the original data.

We obtained six clusters by our proposed model and the BIC of this model is 2501.473 Figure 13 shows the well clustered membership.

4.5. Discussion

We have proposed a Bayesian model for curve clustering and identified genes which has similar trajectory function over time. We have used a Bayesian hierarchical model and Dirichlet process prior of discrete wavelet coefficients. And by product of it, we obtained clustering result in each iteration and we used the marginal posterior mode of the clustering membership of genes. Additionally, we easily estimated the missing data using the conditional distribution. Finally, we finish this chapter by

pointing out two potential shortcomings of the proposed procedure. First, it should be remarkedthat the procedure of this chapter can not be applied to classification problem, known the number of clusters. Second, if the time is not equally spaced, we could not use the discrete wavelet transform directly, however, the latter can be handled by a lifting technique.

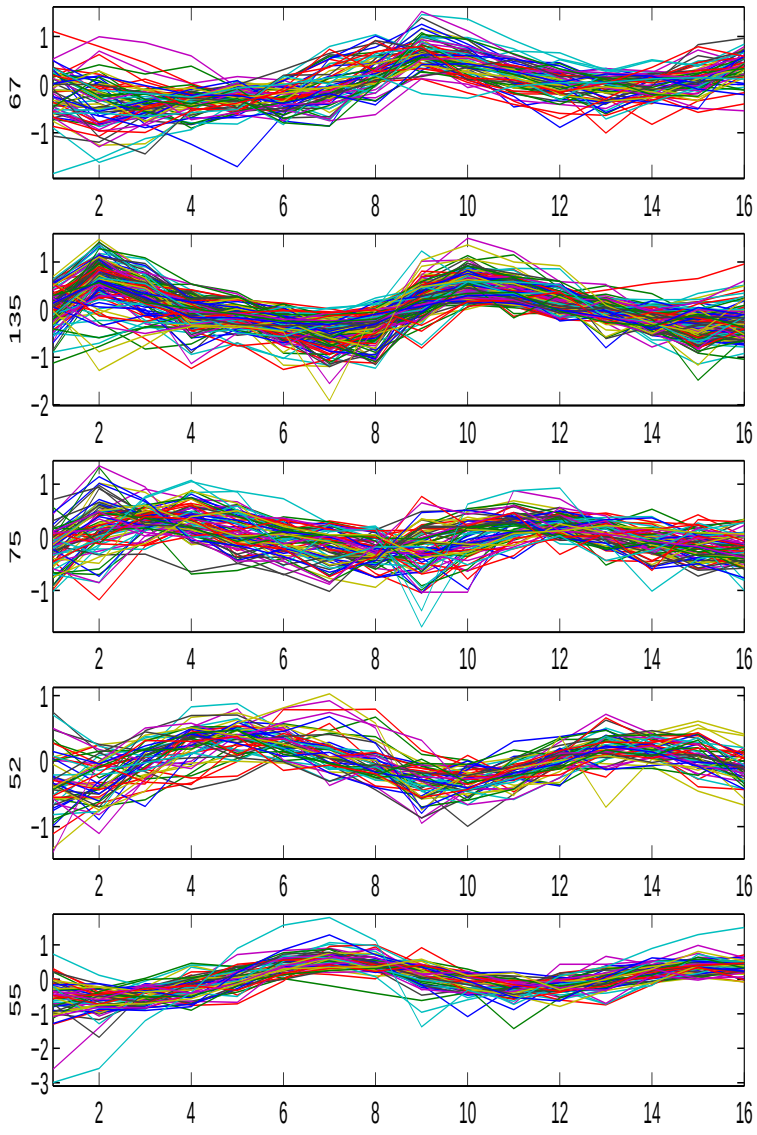

Fig. 10. Five Clusters of Expression Time Courses in Yeast Data (Cho *et al.*, 1998)

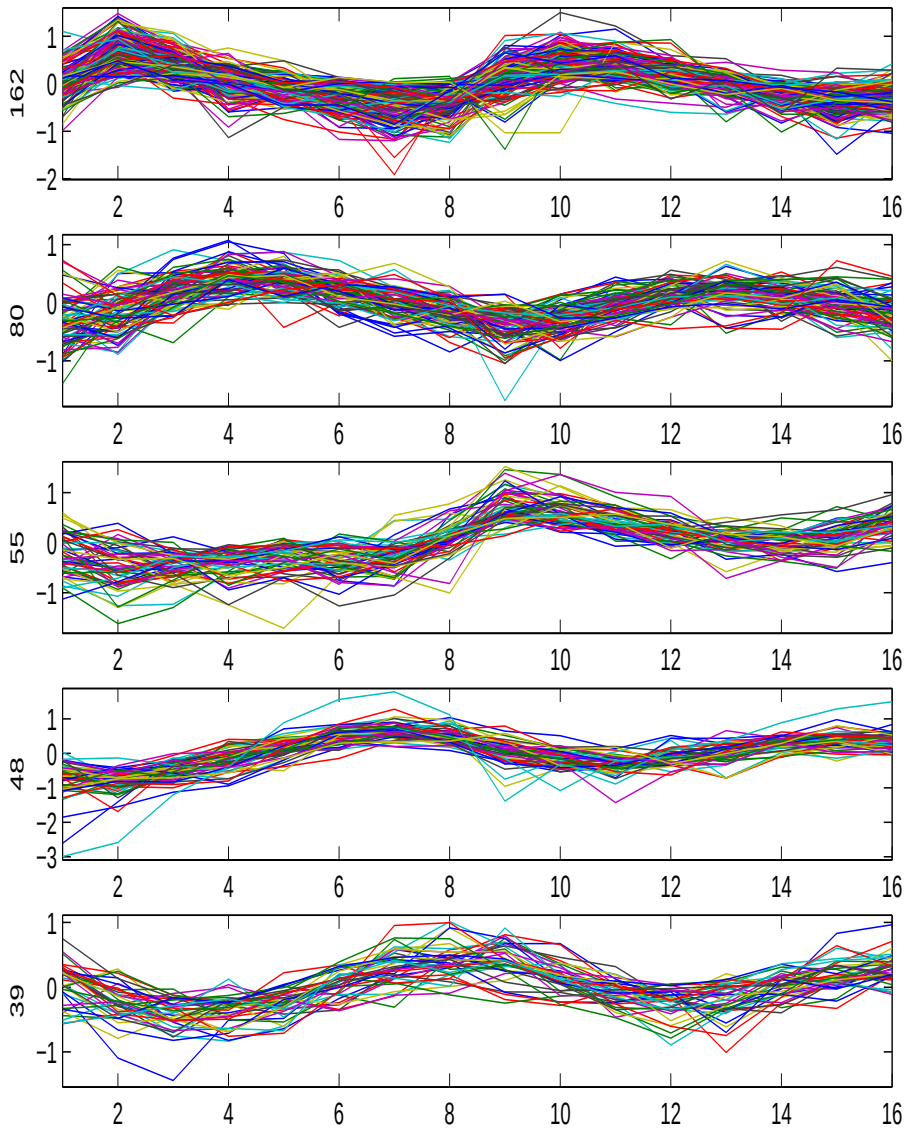

Fig. 11. Five Clusters of Expression Time Courses by Our Proposed Model in Yeast Data (Cho *et al.*, 1998)

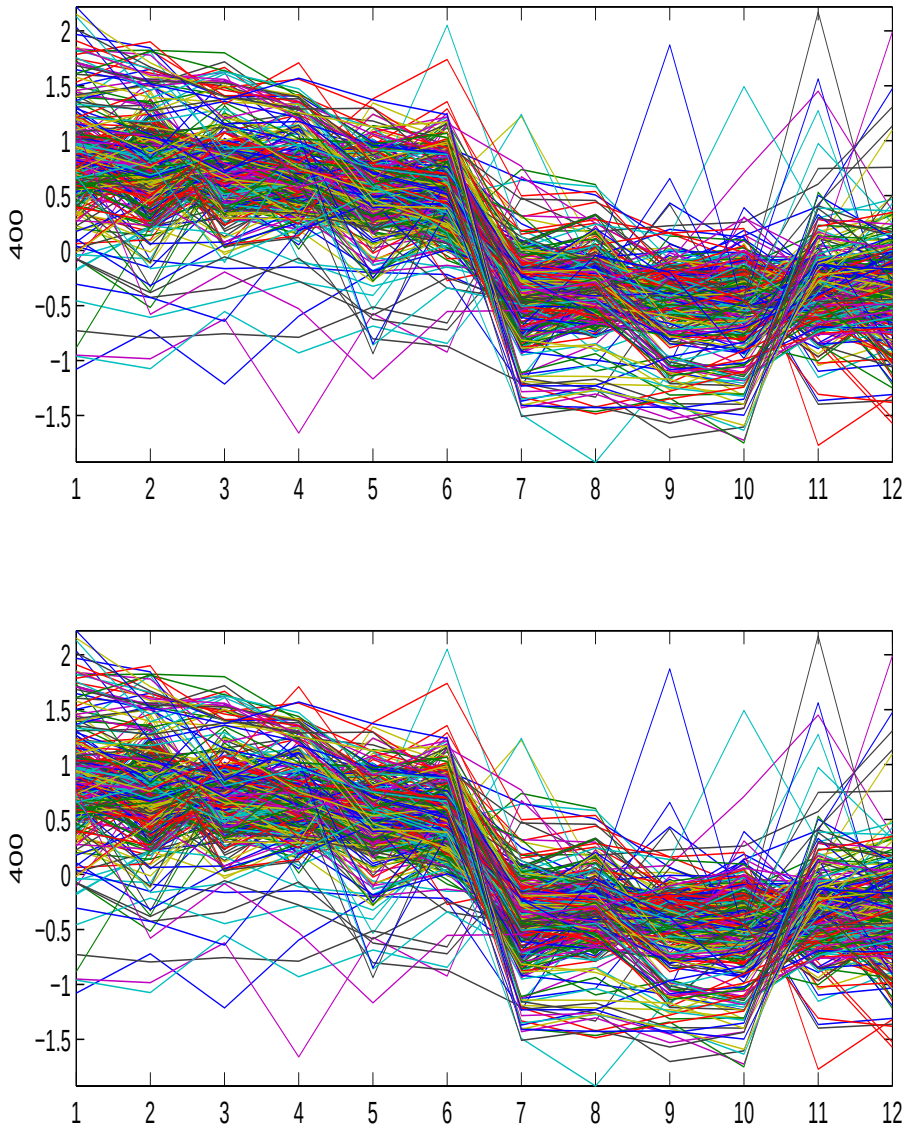

Fig. 12. Original Data versus Estimated Data in Yeast Data (Spellman *et al.*, 1998)

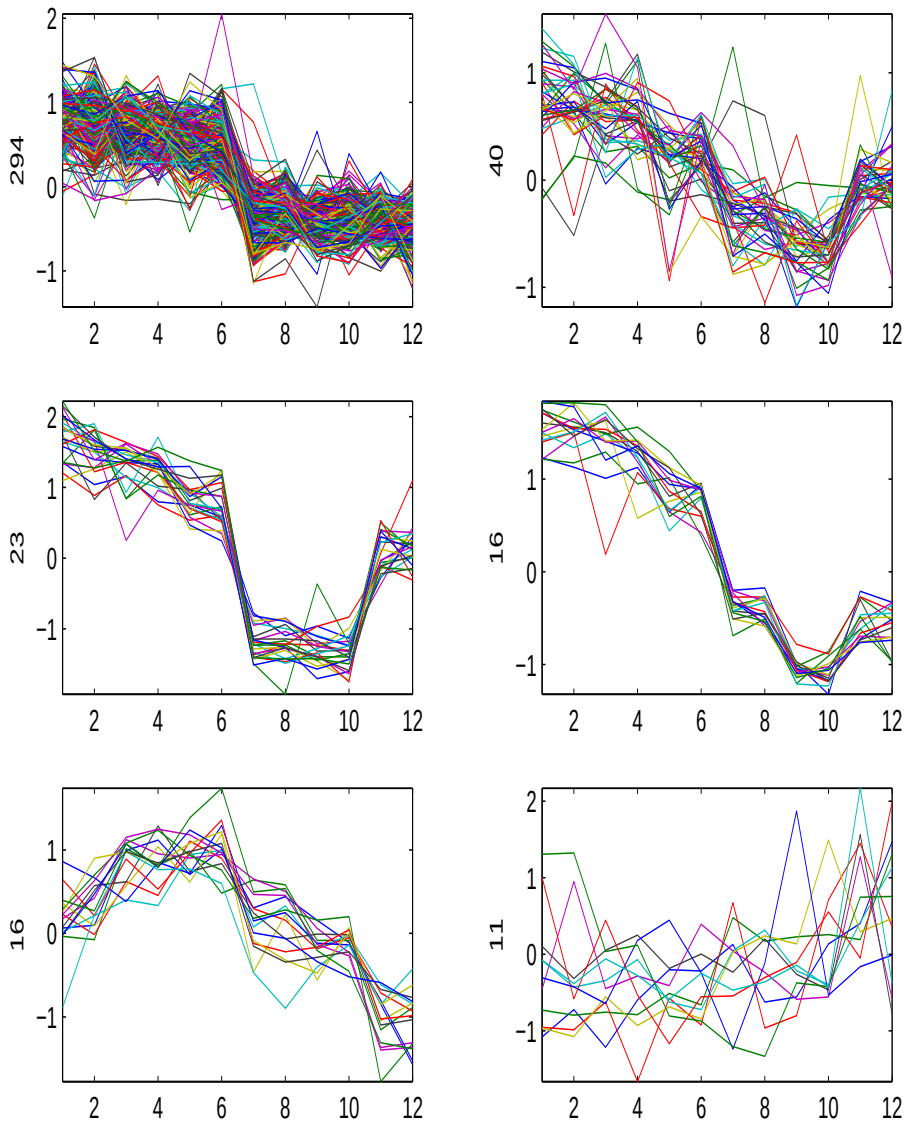

Fig. 13. Six Clusters of Expression Time Courses by Our Proposed Model in Yeast Data (Spellman *et al.*, 1998)

CHAPTER V

SUMMARY AND DISCUSSION

The objective of this research is to propose new Bayesian approaches in variable selection and clustering for application to the cDNA microarray data. At first, we have introduced some background of microarray and reviewed some related literatures. In Chapter II, we have proposed a Bayesian hierarchical model to identify important genes using expression level data to characterize binary categories of samples. We have used a hierarchical probit model and MCMC based stochastic search technique to obtain the posterior samples. We have mainly used Gibbs algorithm in this paper but in future we will investigate the more adaptive MH algorithm (or a mixture of two) to speed up our computation.

Here we have fixed the π value but we can extend our model assuming π is an unknown model parameter. Assigning a conjugate beta distribution prior on π, the extension is straightforward.

Though we have assumed the genes are independent but in our framework we can very easy extend it for dependent cases. For example: suppose if the event that the ith gene is expressed increases the chance that jth gene will be expressed too. In our frame work we can include it through the prior distribution of γ. Rather than taking all the γ_i are independently distributed we can use a Markov model whose transition matrices will be defined as $p(\gamma_j = 1|\gamma_i = 1)$ or so. This type of problems will be handled in future research.

We have considered binary data. Extension to more than two categories can be found in Albert and Chib (1993) and development of a variable selection model in that setup is in Sha (2002).

In Chapter III, we have proposed Bayesian models for variable selection in the

Weibull survival regression model and Cox's proportional hazard model with specific application to analyze microarray data. We have obtained nice estimates of the survival curves with an extremely small number of genes. On the other hand, bigger families of genes can be useful to biologists in studying the relationships and functions. Information on the size of models for prediction can be easily included in our Bayesian search of good models. The method has the flexibility of allowing the location of larger sets of genes, via the inspection of the best visited models or the marginal probabilities of single genes, as we have illustrated.

In Chapter IV, we have proposed a new model-based approach for curve clustering. We have proposed a mixture of Dirichlet process model using discrete wavelet transform for a curve clustering as a fully Bayesian approach. In order to characterize these time-course gene expresssions, we have considered them as trajectory functions of time and gene-specific parameters and obtained their wavelet coefficients by discrete wavelet transform. Then we have clustered curves based on them useing a mixture of Dirichlet process. Each iteration of an MCMC algorithm generates the cluster structure of these coefficients as a by–product (Escobar *et al.,* 1998). We have used a marginal posterior mode of their cluster memberships.

REFERENCES

Albert, J. and Chib, S.(1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88**, 669–679.

Alizadeh, A., Eisen M.B., Davis, R.E., Ma, C., Lossos, I.S. *et al.* (2000). Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling. *Nature* **403**. 503–511.

Alon, U, Barkai, N., Notterman, D.A., Gish, K., Ybarra, S. *et al.* (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745–6750.

Antoniak, C.E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics* **2**, 1152–1174.

Chib, S. and Jeliazkov, I. (2001). Marginal Likelihood From the Metropolis-Hastings Output. *Journal of the American Statistical Association* **96**, 270–281.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I.(1998). The Transcriptional Program of Sporulation in Budding Yeast. *Science* **282**, 699–705.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2nd edition. Boca Raton: Chapman & Hall/CRC.

Cox, D.R. (1972). Regression Models and Life Tables. *J. R. Statist. Soc. B* **34**, 187–220.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia: SIAM.

Decraene, C., Brugg,B., Ruberg, M., Eveno, Matingou, C., Tahi, F., Mariani,

J.,Auffray, C. and Pietu, G. (2002). Identification of Genes Involved in Ceramide-Dependent Neuronal Apoptosis Using cDNA Arrays, *Genome Biol.* **3**(8), research0042.1-research0042.22

Denison, D, Holmes, C., Mallick, B. and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. London: Wiley.

Dudoit, Y, Yang, H, Callow. M and Speed, T. (2000). Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Technical report #578, Stanford: Stanford University.

Duggan, D.J (1999). Expression Profiling Using cDNA Micrarrays. *Nature Genetics* **21**, 10–14.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster Analysis and Display of Genome–Wide Expression Patterns. *Proc. Natl. Acad. Sci. USA***95**, 14863–14868.

Escobar, M.D. (1998). The Effect of the Prior on Nonparametric Bayesian Methods. Manuscript, Toronto: University of Toronto.

Escobar, M.D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Escobar, M.D. and West, M. (1998). Computing Nonparametric Hierarchical Models. P. Muller D.D. Dey and D. Sinha, editors, In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics no. 133, New York: Springer-Verlag.

Ferguson, T.S. (1974). Prior Distributions on Spaces of Probability Measures. *Annals of Statistics* **2**, 615–629.

Friend, S.H. and Stoughton, R.B. (2002). The Magic of Microarrays. *Scientific American* **February**, 44–49.

Gelfand, A. and Smith, A. F. M. (1990). Sampling–Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**, 398–409.

George, E and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* **88**, 881–889.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M. *et al.* (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286**, 531–537.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P.O., Ross, D. *et al.* (2000). Gene Shaving: a New Class of Clustering Methods for Expression Arrays. Technical Report, Stanford: Department of Statistics, Stanford University.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M. *et al.* (2001). Gene Expression Profiles in Hereditary Breast Cancer. *The New England Journal of Medicine* **344**, 539–548.

Hubert, L., and Arabie, P.(1985). Comparing Partitions. *Journal of Classification* **2**, 193–218.

Ibrahim, J.G. and Chen, M.H. (2000). Bayesian Methods for Variable Selection in the Cox Model. In *Generalized Linear Models: A Bayesian Perspective*. Edited by Dipak, K.D., Sujit, K.G. and Mallick, B.K. , New York: Marcel Dekker, 287–309.

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer–Verlag.

Kalbfleisch, J.D. (1978). Non-parametric Bayesian Analysis of Survival Time Data. *J. R. Statist. Soc. B* **40**, 214–221.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data.* 2nd edition. Wiley Series in Probability and Statistics. New York: John Wiley & Sons.

Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation for Incomplete Observations. *Journal of American Statistical Association* **53**, 457–481.

Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S.B. *et al.* (1998). Gene Expression Profiling of Alveolar Rhabdomyosarcoma with cDNA Microarrays. *Cancer Research* **58**, 5009–5013.

Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M. *et al.* (2001). Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks. *Nature Medecine* **7**, 673–679.

Kim, S., Dougherty, E.R., Barrera, J., Chen, Y., Bittner, M.L. and Trent, J.M. (2001). Strong Feature Sets from Small Samples. *Journal of Computational Biology* **9** 129–148.

Lander, E.S. (1999). Array of Hope. *Nature* **21** Supplement 3–4.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes Estimates for the Linear Models (with discussion). *Journal of Royal Statistical Society* **34**, 1–41.

Luan, Y. and Li, H. (2003). Clustering of Time–Course Gene Expression Data Using a Mixed–Effects Model with B–Splines. *Bioinformatics* **19**, 474–482.

MacEachern, S.N., and Muller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Model.* London: Chapman and Hall.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21**, 1087–91.

Milligan, G.W., Soon, S.C. and Sokol, L.M. (1983). The Effect of Cluster Size, Dimensionality, and the Number of Clusters on Recovery of True Cluster Structure *IEEE Trans. Pattern Anal. Mach. Intell.* **5**, 40–47

Nguyen, D.V., Arpat, A.B., Wang, N. and Carroll, R.J. (2002). DNA Microarray Experiments: Biological and Technological Aspects. *Biometrics* **58**, 701–717.

Raftery, A.E., Madigan, D. and Volinsky, C.T. (1995). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance. In *Bayesian Statistics* , volume 5 (Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, 323–350.

Ramesh., N, Anton, I.M., Hartwig, J.H, Geha, R.S. (1997). WIP, a Protein Associated with Wiskott–Aldrich Syndrome Protein, Induces Actin Polymerization and Redistribution in Lymphoid Cells. *Proc. Natl. Acad. Sci* **94**(26) 14671–14676.

Rand, W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846–850.

Rimokh, R., Gadoux, M., Bertheas, M.F., Berger, F., Garoscio, M., Deleage, G., Germain, D., Magaud, J.P. (1993). FVT-1, a Novel Human Transcription Unit Affected by Variant Translocation t(2;18)(p11;q21) of Follicular Lymphoma. *Blood*

81, 136–142.

Robert, C. (1995). Simulation of Truncated Normal Variables. *Statistics and Computing* **5**, 121–125.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**, 467–470.

Schliep, A., Schönhuth, A. and Steinhoff, C. (2003). Using Hidden Markov Models to Analyze Gene Expression Time Course Data. *Bioinformatics* **19**, 255–263.

Sha, N. (2002).*Bolstering CART and Bayesian Variable Selection Methods for Classification* . Ph.D. dissertation, Department of Statistics, Texas A&M University.

Smith, M. and Kohn, R. (1997). Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics* **75**, 317–344.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S. *et al.* (2001) Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998). Comprehensive Identification of Cell Cycle–Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999). Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912.

Theillet, C., Adelaide, J., Louason, G., Bonnet-Dorion, F., Jacquemier, J. *et al.* (1993). FGFRI and PLAT Genes and DNA Amplification at 8p12 in Breast and Ovarian Cancers. *Genes Chromosomes Cancer* **7**, 219–226.

Wakefield, J., Zhou, C. and Self, S. (2003). Modelling Gene Expression over Time: Curve Clustering with Informative Prior Distributions. *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting*, Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M. and West, M. (editors), Oxford: Oxford University Press.

West, M., Nevins, J.R., Marks, J.R., Spang, R. and Zuzan, H. (2000). Bayesian Regression Analysis in the "Large p, small n" Paradigm with Application in DNA Micrarray Studies. Technical Report, Durham: Duke University.

Zheng, W., Xie, DW., Jin, F., Cheng, J.R., Dai, Q., Wen, W.Q., Shu, X.O., Gao YT. (2000). Genetic Polymorphism of Cytochrome P450-1B1 and Risk of Breast Cancer *Cancer Epidemiol Biomarkers Prev.* **9**(2):147–150.

APPENDIX A

DERIVATION OF THE COMPLETE CONDITIONAL DISTRIBUTIONS

$$\begin{aligned} Z|\beta_\gamma &\sim N(X_\gamma\beta_\gamma, 1) \\ \beta_\gamma &\sim N(0, c(X'_\gamma X_\gamma)^{-1}) \end{aligned}$$

$$\begin{aligned} p(Z|\beta_\gamma)p(\beta_\gamma) &\propto \exp\{-\frac{1}{2}(Z - X_\gamma\beta_\gamma)'(Z - X_\gamma\beta_\gamma)\} \\ &\times \frac{1}{c^{q_\gamma/2}|X'_\gamma X_\gamma|^{-1/2}} \exp\{-\frac{1}{2c}\beta'_\gamma(X'_\gamma X_\gamma)\beta_\gamma\} \end{aligned}$$

where $q_\gamma = \sum \gamma_i$.

Since

$$\begin{aligned} p(Z|\beta_\gamma)p(\beta_\gamma) &\propto \frac{1}{c^{q_\gamma/2}|X'_\gamma X_\gamma|^{-1/2}} \exp\{-\frac{1}{2}(1 + c^{-1})\beta'_\gamma(X'_\gamma X_\gamma)\beta_\gamma + Z'X_\gamma\beta_\gamma\} \exp\{-\frac{1}{2}Z'Z\} \\ &= \exp\{-\frac{1}{2}\beta'_\gamma V_\gamma^{-1}\beta_\gamma + \beta'_0 V_\gamma^{-1}\beta_\gamma - \frac{1}{2}\beta'_0 V_\gamma^{-1}\beta_0\} \frac{1}{c^{q_\gamma/2}|X'_\gamma X_\gamma|^{-1/2}} \exp\{-\frac{1}{2}Z'Z\} \end{aligned}$$

where

$$V_\gamma = (1 + c^{-1})^{-1}(X'_\gamma X_\gamma)^{-1},$$

$$\beta_0 = V_\gamma X'_\gamma Z = (1 + c^{-1})^{-1}(X'_\gamma X_\gamma)^{-1}X'_\gamma Z,$$

and

$$\frac{|V_\gamma|^{1/2}}{(c^{q_\gamma}|X'_\gamma X_\gamma|)^{1/2}} = \Big(\frac{1}{1+c}\Big)^{q_\gamma/2},$$

$$\begin{aligned} p(Z|\gamma) &\propto \int p(Z|\beta_\gamma)p(\beta_\gamma)\mathrm{d}\beta_\gamma \\ &\propto \Big(\frac{1}{1+c}\Big)^{q_\gamma/2} \exp\{-\frac{1}{2}S(\gamma)\}. \end{aligned}$$

where $S(\gamma) = Z'Z - \dfrac{c}{1+c}Z'X_\gamma(X'_\gamma X_\gamma)^{-1}X'_\gamma Z.$

So

$$\begin{aligned} p(\gamma|Z) &\propto p(Z|\gamma)p(\gamma) \\ &\propto (1+c)^{-q_\gamma/2}\exp(-\frac{1}{2}S(\gamma)). \end{aligned}$$

Since

$$\begin{aligned} p(\gamma_i|Z,\gamma_{j\neq i}) &\propto p(Z|\gamma)p(\gamma_i) \\ &\propto \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}(1+c)^{-q_\gamma/2}\exp(-\frac{1}{2}S(\gamma)), \\ p(\gamma_i=1|Z,\gamma_{j\neq i}) &\propto \pi_i(1+c)^{-q_{\gamma^1}/2}\exp(-\frac{1}{2}S(\gamma^1)) \\ p(\gamma_i=0|Z,\gamma_{j\neq i}) &\propto (1-\pi_i)(1+c)^{-q_{\gamma^0}/2}\exp(-\frac{1}{2}S(\gamma^0)) \end{aligned}$$

where $\gamma^1=(\gamma_1,\cdots,\gamma_i=1,\cdots,\gamma_p)$ and $\gamma^0=(\gamma_1,\cdots,\gamma_i=0,\cdots,\gamma_p)$,

$$p(\gamma_i=1|Z,\gamma_{j\neq i})=\frac{1}{1+\frac{1-\pi_i}{\pi_i}(1+c)^{1/2}\exp\{-\frac{1}{2}(S(\gamma^0)-S(\gamma^1))\}}$$

Since $p(\beta_\gamma|Z)\propto p(Z|\beta_\gamma)p(\beta_\gamma)$,

$$p(\beta_\gamma|z)\propto\exp\{-\frac{1}{2}(\beta_\gamma-\beta_0)'V_\gamma^{-1}(\beta_\gamma-\beta_0)\}.$$

So the posterior distribution of β_γ is

$$\beta_\gamma|Z\sim N(\beta_0,V_\gamma)$$

VITA

Kyeong Eun Lee was born in Taegu, Korea on December 14, 1971. She is the first daughter of Tae-dong Lee and Kae-Soon Lee. She graduated from Kyungpook National University in Taegu, Korea in August 1993 with a Bachelor of Science degree in statistics. In February 1997, she received a Master of Science degree in statistics working under Dr. Byeong Uk Park from Seoul National University in Seoul, Korea. She moved to America to pursue her Ph.D. in August of 1998 in Department of Statistics of Texas A&M University. In May of 2004 she received her Ph.D.

Her permanent address is

897-22 Pum-A 1 Dong Su-Sung Gu

Taegu, Republic of Korea